30820

NATIONAL LIBRARY
OTTAWA

CANADA

BIBLIOTHÈQUE NATIONALE
OTTAWA

NAME OF AUTHOR..... Edward H. Scissons

TITLE OF THESIS..... Convergence of Clinical Judgement: A Multitrait Analysis

UNIVERSITY..... Alberta

DEGREE FOR WHICH THESIS WAS PRESENTED..... Ph.D.

YEAR THIS DEGREE GRANTED..... Fall 1976

Permission is hereby granted to THE NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(Signed)..... [signature]

PERMANENT ADDRESS:

519 Ave G So

Saskatoon, Sk

.....

DATED..... June 14 1976

NL-91 (10-68)

INFORMATION TO USERS

THIS DISSERTATION HAS BEEN
MICROFILMED EXACTLY AS RECEIVED

This copy was produced from a microfiche copy of the original document. The quality of the copy is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

PLEASE NOTE: Some pages may have indistinct print. Filmed as received.

Canadian Theses Division
Cataloguing Branch
National Library of Canada
Ottawa, Canada K1A ON4

AVIS AUX USAGERS

LA THESE A ETE MICROFILMEE
TELLE QUE NOUS L'AVONS RECUE

Cette copie a été faite à partir d'une microfiche du document original. La qualité de la copie dépend grandement de la qualité de la thèse soumise pour le microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

NOTA BENE: La qualité d'impression de certaines pages peut laisser à désirer. Microfilmee telle que nous l'avons reçue.

Division des thèses canadiennes
Direction du catalogage
Bibliothèque nationale du Canada
Ottawa, Canada K1A ON4

THE UNIVERSITY OF ALBERTA

CONVERGENCE OF CLINICAL JUDGEMENT: A MULTITRAIT ANALYSIS

by

EDWARD H. SCISSONS

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

EDMONTON, ALBERTA

FALL, 1976

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled Convergence of Clinical Judgement: A Multitrait Analysis submitted by Edward H. Scissons in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

Supervisor

External Examiner

Date May 27, 1976

ABSTRACT

This research was a study of the convergence of clinical judgement (multitrait ratings) across three different information sources (psychometric tests, interview, and test + interview). Of major interest was the similarity of clinical evaluations of ability across the three different information sources.

Subjects (N=74) were executives appraised by a firm of industrial psychologists. Subjects were evaluated independently on 18 different traits on the basis of: test information alone, interview information alone, or test and interview information combined.

Results indicate a varying degree of convergence of clinical ratings dependent on clinician and trait. A clinician by factor by rating condition model of executive assessment is developed. Convergence indices ranged from a high of .64 to a low of .05. The nature of reliability theory, as it pertains to clinical judgement research, is discussed and suggestions for further research in the area are presented.

ACKNOWLEDGEMENTS

I would like to acknowledge the contribution of many friends who assisted in the development of this research project.

The clinicians and staff of A. W. Fraser & Associates (Al Fraser, Jim Wuest, John Roshak, Pat Mitchell, and Libby Bolstler) gave unconditionally of their time and support. Without them, this project would not have been possible.

I am also indebted to the members of my committee: Dr. George Fitzsimmons (Chairman), Dr. Charles Anderson, Dr. Harvey Zingle, Dr. Murray Smith, and Dr. Lloyd Njaa (External). Their advice and counsel was appreciated.

Most importantly, I would like to thank Linda and Paddy who were forced to understand the neurosis that is research.

Financial support during the course of this research was provided by a University of Alberta Dissertation Fellowship.

Ed Scissons

TABLE OF CONTENTS

CHAPTER		PAGE
I.	THE PROBLEM	1
II.	REVIEW OF THE LITERATURE	5
	Models of Clinical Judgement	6
	Validity and Reliability of Clinical Judgement	9
	Clinical Judgement in Executive Appraisal	16
	Summary	21
III.	EXPERIMENTAL DESIGN	23
	Procedure	24
	Experimental Hypotheses	26
	Limitations of the Study	26
IV.	RESULTS	28
	Definition of Terms	28
	Inter-Rater Reliability: Test Condition	92
	Factor Analysis of Test Condition Ratings	95
	Summary	98
V.	DISCUSSION	99
	Interpretive Problems	105
	Conclusions and Implications	109
	Suggestions for Further Research	111
REFERENCES		114
APPENDIX 1.	DEFINITION OF CHARACTERISTICS	121
APPENDIX 2.	INTERVIEW RATING FORM	125
APPENDIX 3.	INTERVIEW + TEST RATING FORM	127

APPENDIX 4. TEST RATING FORM 129

APPENDIX 5. DUNNETTE (1971) TABLE 1 131

APPENDIX 6. FACTOR ANALYSIS OF 18 CHARACTERISTICS 133

APPENDIX 7. CAPSULE SUMMARY OF TESTS 137

LIST OF TABLES

Table	Description	Page
1	ANOVA: Factor 1	33
2	Reliability: Factor 1	34
3	Means and Standard Deviations: Factor 1	34
4	ANOVA: Factor 2	37
5	Reliability: Factor 2	38
6	Means and Standard Deviations: Factor 2	38
7	ANOVA: Factor 3	41
8	Reliability: Factor 3	42
9	Means and Standard Deviations: Factor 3	42
10	ANOVA: Factor 4	44
11	Reliability: Factor 4	45
12	Means and Standard Deviations: Factor 4	45
13	ANOVA: Factor 5	48
14	Reliability: Factor 5	49
15	Means and Standard Deviations: Factor 5	49
16	ANOVA: Factor 6	51
17	Reliability: Factor 6	52
18	Means and Standard Deviations: Factor 6	52
19	ANOVA: Factor 7	54
20	Reliability: Factor 7	55
21	Means and Standard Deviations: Factor 7	55
22	ANOVA: Factor 8	58
23	Reliability: Factor 8	59
24	Means and Standard Deviations: Factor 8	59

25	ANOVA: Factor 9	61
26	Reliability: Factor 9	62
27	Means and Standard Deviations: Factor 9	62
28	ANOVA: Factor 10	65
29	Reliability: Factor 10	66
30	Means and Standard Deviations: Factor 10	66
31	ANOVA: Factor 11	68
32	Reliability: Factor 11	69
33	Means and Standard Deviations: Factor 11	69
34	ANOVA: Factor 12	71
35	Reliability: Factor 12	72
36	Means and Standard Deviations: Factor 12	72
37	ANOVA: Factor 13	74
38	Reliability: Factor 13	74
39	Means and Standard Deviations: Factor 13	75
40	ANOVA: Factor 14	77
41	Reliability: Factor 14	78
42	Means and Standard Deviations: Factor 14	78
43	ANOVA: Factor 15	80
44	Reliability: Factor 15	81
45	Means and Standard Deviations: Factor 15	81
46	ANOVA: Factor 16	83
47	Reliability: Factor 16	84
48	Means and Standard Deviations: Factor 16	84
49	ANOVA: Factor 17	87
50	Reliability: Factor 17	88

51	Means and Standard Deviations: Factor 17	88
52	ANOVA: Factor 18	90
53	Reliability: Factor 18	91
54	Means and Standard Deviations: Factor 18	91
55	Inter-Rater Reliability Estimates	94

CHAPTER I

STATEMENT OF THE PROBLEM

"Progress in psychological assessment is important not only to such applied fields as clinical, counselling, educational, and industrial psychology, but is vital also to the continued development of psychology as-a-whole (McReynolds, 1968, p. 1)". Modern psychology is directly concerned with understanding the human condition; research in psychology has traditionally been oriented towards the development and evaluation of better assessment techniques and procedures (McReynolds, 1968).

The "clinical judgement debate", as it has been called, developed as a vigorous movement in psychology in the early 1950's. Of concern were problems such as the ability of psychologists to predict future behavior, the validity and reliability of prediction, and clinical versus actuarial methods of prediction. Early writings, such as that of Meehl (1954), did much to spark debate between the psychodiagnosticians on one hand and the actuarially oriented researcher or clinician on the other. Of most importance was the accuracy (in all senses of the word) of assessment decisions based on either clinical or actuarial integration of client information. The controversy is far from over, but research of late has concentrated more on improving both methods of prediction or decision making rather than fanning the fires of difference that exist between the two (Goldberg, 1970).

Managers, administrators, and other executives play an important role in modern society and are always in short supply

(Dunnette, 1971). This is not to imply that managers and other professionals are in short supply but rather that good managers, good administrators, and good executives remain a scarce commodity in the occupational marketplace.

Industrial psychologists and other professionals concerned with what makes a good executive and how to identify a good executive by methods other than trial and error, are involved in a specialized aspect of the clinical judgement dilemma. Research here has focused on studies concerned with the predictive validity of executive assessments and studies which investigated the assessment process itself from the points of view of validity and reliability. Thus we have studies such as that by Bray & Grant (1966) that investigate the specific contribution of the interview to over-all executive assessment and studies such as that by Wollowick & McNamara (1969) which look at the components of an executive assessment program. Other research, not specifically dealing with clinical judgement, has been concerned with the interview as a diagnostic technique (Webster, 1964; Grant & Bray, 1969; Ulrich & Trumbo, 1965; Mayfield, 1964), testing as an adjunctive or sole means of executive assessment (Henrichs, 1969; Spitzer & McNamara, 1964; Bray & Moses, 1972), or various multiple assessment techniques (Albrecht, Glaser & Marks, 1964; Wollowick & McNamara, 1969; Campbell, Otis, Liske & Prien, 1962).

The relationship of clinical judgement to executive appraisal is a logical one. Clinical judgement is concerned with assessment;

those dealing with executive appraisal are also concerned with assessment at a very operational level. Although most research in the area of clinical judgement has been concerned with unidimensional decision making, e.g., the diagnosis of psychotic versus neurotic from MMPI profiles (Goldberg, 1965), some researchers (Goldberg & Werts, 1966; Donaldson, 1969) have addressed themselves to a more complex multitrait multimethod approach (Campbell & Fiske, 1959). There is, however, little research which relates these clinical judgement findings from clinical psychology to the multitrait multimethod domain of executive appraisal. There has been virtually no work, other than that concerned with assessment centers, which relates this multitrait multimethod model to executive appraisal in a natural setting. It is the use of this natural setting which is most likely to result in research findings high in generalizability as a result of high ecological (external) validity (Snow, 1974).

Even within the domain of clinical judgement in clinical psychology, most research has focused on prediction accuracy, stability, or concensus rather than convergence as measures of judgement effectiveness. Convergence in clinical judgement is important because it yields a measure of the degree of similarity in the assessments a clinician makes with respect to his clients as a result of the different types of data available about these clients, e.g., test versus nontest data (Goldberg & Werts, 1966).

This study is an investigation of the convergence of clinical judgement in executive appraisal. The hypothesis tested is that there will be a significant difference in the assessment of a client by a clinician depending on the type of information available about that client. Of specific interest in this study are the differences in appraisal (multitrait ratings) as a result of information obtained by (a) interview alone, (b) testing alone, or (c) testing + interview combined.

This study is of considerable importance at both a theoretical and an operational level. At a theoretical level, rationale for the study focus on the generalizability of clinical findings across data bases, nature of the interaction between trait, information base, and clinical judgement, particularly as these affect multitrait analysis of ability, and the providing of an empirical base for further predictive validity studies once the problem of convergence has been accounted for. At present, there exists no research to provide a rationale for the generalizing of clinical judgement findings across data bases; there appears to be an unmet assumption of high convergence.

At an operational level, this study is important because it is concerned with the possible duplication of psychological services. If high convergence is evident on several or all of the traits involved in this multitrait analysis, cost alone should dictate a judicious duplication of services through multimethod assessment techniques.

CHAPTER II

REVIEW OF THE LITERATURE

"To many people, the prediction problem must seem to be the basic problem of applied psychology (Gough, 1962, p. 526)". Studies of clinical judgement, which are only one aspect of the 'prediction problem' discussed at length by Gough (1962), have progressed through a number of rather distinct stages if viewed in a historical perspective (Bieri, Atkins, Briai, Leaman, Miller, & Tripodi, 1966). Research has developed from its roots in introspective analysis (Erickson, 1959) to studies of the validity and reliability of clinical judgement, clinical versus statistical prediction (Meehl, 1954), and on to the most recent stage which is concerned with models of decision making within the framework of decision theory. In many ways, studies concerned with the validity/ reliability of clinical judgement and those concerned with actuarial versus clinical predictive validity are similar. Both are concerned with improving and/or describing the decision making process directly, i.e., in terms of outcomes. The last stage, model building, has been an attempt to develop theoretical models of decision making or information processing as an indirect attempt to improve future decisions (Bieri et al, 1966) rather than to à priori evaluate present ones.

This literature review will examine clinical judgement from three perspectives: (1) models of clinical judgement, (2) reliability and validity of clinical judgement including the actuarial versus

clinical dilemma, and (3) clinical judgement in executive appraisal.

Models of Clinical Judgement

Since the early 1960's the focus on clinical judgement research has been concerned with the nature of the clinical judgement decision making process itself. Of major concern has been the development of mathematical models to either explain or improve on the actual judgement of the clinician.

Goldberg (1971) isolates two general models for clinical decision making; linear and non-linear. The linear model is that model expressed by a multiple regression analysis and is equivalent to the formulation of regression weights in order to combine accurately available information for purposes of prediction. Non-linear models usually involve some type of moderator variable effect; i.e., the weighting of one variable will vary in relation to the magnitude of the difference between two or more other variables. A number of different types of non-linear models have been postulated (Goldberg, 1971; Einhorn, 1970, 1971) all involving some form of moderator variable combination.

Wiggins & Hoffman (1968) outline an important study which examines the relative efficacy of three different models of information combination; the linear, quadratic, and sign models. The quadratic model is similar to the linear model already described but includes the squares and products of the original linear model. The sign model incorporates a linear combination of 70 clinical signs in relation to MMPI interpretation first described by

Goldberg (1965). Their experiment involved an experimental design now classic in clinical judgement research. Psychologists were required to rate MMPI profiles as psychotic or neurotic in a blind rating fashion. Results indicated the presence of both linear and configural processing of information by clinicians dependent on both clinician and subject samples. Clinicians obtained results which were similar to computer integration of information as per the three models just described. However, as noted by the authors, the differences between the results obtained by any of the three methods of information combination were not great. The simple linear model combined data in a very efficacious manner.

Goldberg's (1965) study is further supported by Dawes (1972) and Dawes & Corrigan (1974) who describe two different types of linear models in an experiment designed to test the ability of human judges to perform against even random linear models. The two models, actuarial (based on a regression of the criterion in the predictors) and bootstrapping (based on a regression of the judges' prediction on the predictors) were both superior to the decisions of human judges even when regression weights were assigned randomly rather than systematically. Experiments cited by the two authors involved the rating of psychotic versus neurotic on the MMPI, prediction of graduate school success, and geometric design estimation. Dawes (1972) summarizes his findings: "If a reasonable sample of cases exists for which the output values are known, the best way to make the predictions is to estimate beta weights for

the input variables on the basis of multiple regression; human judges should be ignored (p. 3)".

Wainer (1976) further reinforces this finding. He indicates that, in very general circumstances, little is lost in terms of the original data if regression coefficients are estimated rather than calculated.

Configural processing, best described as a usage of moderator variables either overtly or covertly, has also commanded considerable attention in the clinical judgement literature. Hoffman, Slovic, & Rorer (1968) utilized an ANOVA technique to assess configural processing in the diagnosis of malignant gastric ulcers using nine radiologists as clinicians. Although the authors were able to demonstrate conclusively the reality of configural processing, they further indicate that even when this processing was utilized by the clinicians, clinician decision accuracy did not match even that of a simple linear combinative model.

Einhorn (1972), in an important study involving the clinical judgement of malignant cancers, addresses himself to the efficacy of combining components of the decision making process rather than the binary decisions involved in many of the classic studies in the area. He suggests the use of expert clinicians, in their specific areas of expertise, and combining these 'mini-decisions' mechanically.

Shinedling, Howell, & Carlson (1975) combine both clinical 'rule of thumb' techniques with statistics to produce a 'clinistics' model of clinical judgement. They conclude that, "rather than

trying to justify the utility of personal, private judgement, psychologists should study the contribution of objective clinical decision-making strategies. Studying 'clinistics' might lead to new insights and understandings about behavior (p. 389)".

Goldberg (1970) may be getting much closer to the truth when he describes his very important study which once again utilizes the clinical task of distinguishing psychotic versus neurotic MMPI profiles. He concludes that the model that the clinicians actually used, when applied systematically and consistently, yielded better decisions than did the actual clinicians. The problem with clinicians he argues, may not be that they are wrong, but that they are inconsistent (or human!).

Slovic, Rorer, & Hoffman (1971) carry Goldberg's (1971) research one step further. They investigated the reasons why clinicians diagnose differentially. In a study involving the diagnosis of gastric ulcer malignancy, they attempted to discover how each clinican used the various clinical signs available to him. Their research enables them to trace differential diagnosis back to a differential use of clinical signs. They cite that the major use of their method is in the opportunity afforded in the 'train-to-model' teaching of student clinicians.

The Validity and Reliability of Clinical Judgement

Outcome studies in the area of clinical judgement have focused most directly on the predictive validity of clinical decision making be those decisions made clinically or actuarially.

Meehl (1954), in his now classic book, Clinical versus Statistical Prediction, analyzed previously published studies dealing with the validity and reliability (consistency and stability) of clinical and actuarial decisions. He summarizes his findings:

> In spite of the defects and ambiguities present, let me emphasize the brute fact that we have here, depending upon one's standards for admission as relevant, from 16 to 20 studies involving a comparison of clinical and actuarial methods, in all but one of which the predictions made actuarially were either approximately equal or superior to those made by the clinician. (p. 119)

Although attempting to maintain a balanced perspective in analyzing the clinical versus actuarial dilemma, Meehl (1954) finds himself unavoidably drawn to the side of the actuary. The clinician cannot predict at a level that would rival even the most simple linear regression equation. Meehl (1954) has been taken to task by several other writers because of his handling of the clinical versus actuarial problem.

Holt (1958) rejects as artificial the dichotomy employed by Meehl (1954) of clinicians on one hand and actuaries on the other. He indicates that clinical judgement must enter the actuarial process at frequent intervals. The actuary must still select his tests, criterion measures, intervening variables, and psychological constructs. How then, Holt (1958) argues, can we even talk of such a false distinction. Both are merely forms of clinical integration. In a later treatise, Holt (1970) reaffirms his argument while concluding that the largely actuarial model does have some place in

combining largely numerical information for purposes of decision making.

Sawyer (1966) also sees the problem of clinical versus actuarial decision making as merely the last half of the problem. He indicates that the collection of data can also be considered as a clinical or actuarial problem (e.g., the choice to collect test or interview data). Sawyer (1966) concludes that the real strength of the clinician is in the providing of additional nonpsychometric information to the decision making process and not in decision making per se. Sawyer (1966) indirectly discounts much of the research reviewed by Meehl (1954) by indicating that the paucity of research favoring the clinical method derives from the fact that the research design utilized in many studies has forced the clinician to play the actuarial game (e.g., forced choice responses for ease of tabulation or the exclusion of nonpsychometric information--interview impressions). Holt (1970) reaffirms this view; he says that studies have yet to look at clinical prediction at its best compared with actuarial prediction at its best.

Meehl (1954) describes four combinations of data and methods of obtaining data as (a) psychometric data combined mechanically, (b) psychometric data combined nonmechanically, (c) nonpsychometric data combined mechanically, or (d) nonpsychometric data combined nonmechanically. More complex combinations of these singular combinations are also possible (e.g., psychometric and nonpsychometric data combined nonmechanically). However, the bulk of research that

Meehl (1954) reviews would fall into categories (a) and (b); little evidence is available regarding the more methodologically difficult categories or combinations of categories. It seems that Meehl (1954) is reviewing studies high in experimental rigor but low in ecological validity.

Holt (1970), in his review of Meehl (1954), Holt (1958), Sawyer (1966), and more recent clinical and actuarial findings, concludes that:

> (a) When the necessary conditions for setting up a pure actuarial system exist, the odds are heavy that it can out-perform clinicians judgements in predicting almost anything in the long run if both sides have access only to quantitative data such as an MMPI profile. (b) A complete six-step predictive system is almost always better than a more primitive one, and even when it seems to be entirely statistical, it requires the exercise of a great deal of subjective judgement to work efficiently. (c) Disciplined, analytical judgement is generally better than global, diffuse judgement, but it is not any less clinical. (d) To predict almost any kind of behavior or behavioral outcome, one does better to assess the situation in which the behavior occurs in addition to assessing the actors' personalities. (e) Granted such knowledge and a meaningful criterion to predict, clinical psychologists vary considerably in their ability to do the job, but the best of them can do very well. That is they do have the skills in assessing personality by largely subjective, but partly objecti - fiable procedures, making use of theories that permit a deeper and more valid understanding of persons than anything a statistician can provide. (p. 348)

The real problem of the predictive validity of clinical or actuarial judgement may be escaping both clinician and actuary. Ash & Kroeker (1975) review the efficacy of both models of decision making. They would rate both as low indicating that a criterion-predictor match of .60 (high by today's standards for either clinicial or

actuarial techniques) is still appallingly low.

Clinical Judgement: Reliability

In comparison to both model building and predictive validity studies, a much more limited amount of research has focused on the problems associated with the reliability of clinical judgements. Goldberg & Werts (1966) cite several types of reliability measures of interest to clinical judgement researchers: "(a) over time for the same judges using the same data (stability), (b) over judges, for the same data from the same occasion (concensus), and (c) over data sources administered on the same occasion and interpreted by the same judge (convergence) (p. 199)". Goldberg & Werts (1966) indicate that problems in any one of these areas or, as is more likely, in combinations of these areas, pose threats to the validity of judgement. They see the error covariation across time, sources, traits, and targets as major limitations in the study of clinical judgement. They indicate that, "no study of the reliability of clinical inferences is ever likely to provide definitive conclusions (p. 200)". Sawyer (1966), in discussing the overriding concern with the validity of clinical judgements, comments that simple comparisons between combinative models do little to explain or improve either method.

The classic study of convergence in clinical judgement was done by Little & Schneidman (1959). These researchers were concerned with the convergence of clinical judgement over certain aspects of a similar data base (psychometric data). Clinicians were required

to rate subjects using a Q-sort technique as either psychotic, neurotic, psychosomatic or normal on the basis of one of the Rorschach, Thematic Apperception Test and Make a Person, MMPI or a combination of several interpretive tests. Their findings, while disheartening for the clinician, are not altogether unexpected. They were unable to find a high degree of convergence across similar aspects of the same data base. The problems in generalizing from the Little and Schneidman (1959) study are manifold. They are dealing with a unidimensional data base (psychometric data), are concerned with unidimensional decision making, and are concerned with a psychologically "unwell" population.

Goldberg & Werts (1966) utilize a specialized form of multitrait multimethod clinicial judgement research. Clinicial psychologists were required to rate psychiatric patients on four categories using one of four data sources (MMPI, Rorschach, Wechsler, or Vocational History). They were unable to find any relationship between the judgements of one clinician working from one information source and those of another clinician working from another data source. This study cannot be considered a real study of convergence in clinical judgement since it is concerned more with agreement across raters (concensus) as it is with agreement across sources (convergence). This study would probably score low in what Snow (1974) would call ecological or external validity. There seems to be a real dissimilarity between experimental tasks and "real" clinician tasks in real assessment situations. Experimental

clinicians were asked to rate subjects in a manner which was probably foreign to them and were then chastized for failing to rate consistently. Sawyer (1966) would see this as a study in which the clinician was made to play the actuarial game. This threat to external validity is further magnified by the confounding of concensus and convergence as reliability measures. How important is it that the ratings of one clinician from one data source agree with those of another clinician using a different data source?

Goldberg (1966), in a study of peace corps selection board procedures, evaluated the stability and convergence (inseparably) of board members' decisions regarding potential applicants. The relationship of board members' individual decisions before and after board discussions of the candidates was analyzed. His findings were that decisions before and after board discussion were highly correlated, being in the order of .80 but that decisions between raters were only moderate, being in the order of .40. The study, although interesting, is difficult to interpret because of the confounding of stability and convergence. In terms of its external validity, however, it must be applauded.

Slovic (1966) indirectly addresses himself to the reliability of clinical judgement, particularly across diverse and multiple information sources. His findings indicate that, in the prediction of intelligence, clinicians used only two or three key predictors even when they were presented with (and believed they used) many. Additional sources of information were used only when conflicting

information was evident in the prime two or three factors. A threat to reliability then, may be the targeting behavior of clinicians in reference to the information they have available. This is further confirmed by Perez (1973) in a study involving the discrimination between different types of criminal test protocols. His research indicates that additional information has little effect on decision accuracy (reliability or validity).

The questions of why and how clinical judgements are unreliable (or reliable) remain largely unanswered in the literature. It is noteworthy that few researchers or studies to date have systematically investigated the problems of reliability, particularly convergence, preferring to further reinforce the wealth of information available in the areas of predictive validity.

Clinical Judgement in Executive Appraisal

Although the relationship of clinical judgement research to the field of executive appraisal is a logical one, the area has been only sparsely researched. Historically, the emphasis taken in the derth of information available that deals with executive appraisal and characteristics of successful executives, has focused on the predictive validity of unimethod (interview) or multimethod (assessment centers) assessment techniques. Thus, we have seen very little of model building, as has been the emphasis in clinical judgement in the areas of clinical psychology, or on reliability, a point in common between the two areas.

Ulrich & Trumbo (1965) present an excellent and detailed summary of the personnel selection interview to which the reader is referred. Their findings indicate that the low predictive validity demonstrated by most assessment interviews may be due to contamination of data or criterion problems. They see the major use of the interview in assessing personal relations and career satisfaction. The lack of sufficient controls on interview research is a concern echoed by Mayfield (1964) and Mayfield & Carlson (1972). All of these researchers agree that the major thrust in interviewing research should be internal, i.e., "studying the decision making process as it operates in the selection interview" (Mayfield & Carlson, 1972, p. 41).

Other studies on the interview have shown low stability of ratings (Vaughn & Reynolds, 1951) and low inter-rater reliability (Schwab & Heneman, 1969) on the basis of informal unstructured interviews. Vaughn & Reynolds (1951) indicate that inter-rater reliability (concensus) increases as a direct function of interview structure. Hollman (1972) explains part of the problem regarding intra- or inter-rater reliability in interviewing, particularly with respect to threats to validity. He indicates that interviewers appear unduly swayed by negative client information obtained during interview and tend to ignore more relevant positive information obtained at the same time. Langdale & Wertz (1973) add that inter-rater reliability increases as a function of interviewer knowledge of the prospective job, adding that unless the interviewer knows

the prospective position thoroughly, inconsistencies are inevitable.

Other researchers, in discussing threats to the reliability of the assessment interview (and indirectly, validity), have focused on other areas. Baskett (1973) indicates that a major concern should be the similarity of interviewer-interviewee attitudes. When these attitudes differ markedly, interviewee ratings suffer. Lipsett (1964) argues for the use of interviewing, saying that much of what we think we have with personnel tests (validity), never really existed.

The literature on interviewing in executive appraisal, while plentiful, does not answer much in relation to clinical judgement. We know only that poor or ineffectual decisions are being made. We have little indication of why or where.

The only area of executive appraisal to which a modified form of the clinical judgement research may be applicable is the assessment center. The assessment center, first commercially used by the American Telephone and Telegraph Company to assess managerial performance and potential (Bray & Grant, 1966), is an adaptation of German psychologists procedures for screening officer candidates (Dunnette, 1971; Blumenfeld, 1971). The assessment center combines performance appraisal techniques, such as the interview, paper-and-pencil tests, in-basket exercises, leaderless groups, and simulation exercises to formulate multitrait ratings of candidates. Traditional clinical judgement findings are not directly applicable here since ratings of several psychologists, managers, or super-

visors, although derived independently, are combined for purposes of final assessment. Dunnette (1971) describes the relationship of assessment center findings to behavioral ratings obtained on the job. Correlations ranged from the low ±.20's to the high ±.70's depending on the trait measured (see Appendix 5).

Bray & Grant (1966) studied an assessment center initiated to appraise future managers for the Bell Telephone System. Their findings indicated that, although all predictors were used for making ratings, considerable inter-rater variability was evident in combining the data. In an aspect of the same study, Grant & Bray (1969) dealt more specifically with the interview information obtained in the assessment center. Their findings indicate that structured interviews are able to yield reliable and valid indicators of future performance.

Wollowick & McNamara (1969) in their research which studies the use of the assessment center with IBM managers, found that adding information received from situational tests increased predictability. These researchers also add weight to the actuarial versus clinical debate by adding that a statistical combination of the assessment center program variables was better than any single subjectively derived overall rating. Henrichs (1969), in dealing with the same subject pool as Wollowick & McNamara (1969), indicates that a careful analysis of employee work records was also highly related to future performance.

Moses (1973), in a more recent study of assessment centers,

reinforces this; he notes increased validity of assessment center predictions as a function of increasing time between prediction and evaluation.

Albrecht, Glaser, and Marks (1964) use a multiple assessment procedure that is really a forerunner of the assessment center approach. They were unable to find significant validity in the procedure using a multitrait multimethod matrix approach, but their research was hampered by methodological shortcomings. Criterion behaviors were evaluated by superiors who had little contact with candidates or by peers rather than by direct supervisors.

Bray & Grant (1966) indicate that many of the key characteristics measured by the assessment center can be obtained by an interview, a finding suggested by Glaser, Schwarz, and Flanagan (1958), but one that is at variance with more disheartening research on the assessment interview (Webster, 1964).

Blumenfeld (1971) sees the greatest benefit in assessment center methodology as the equal opportunity afforded candidates, use of trained assessors, and situational exercises high in what Snow (1974) would call ecological validity. Wilson & Tatge (1973) are less optimistic; they see the assessment center approach as very costly and not necessarily better than more traditional methods of assessment.

Trankell (1959) describes a study which, although it deals almost exclusively with predictive validity, is noteworthy in terms of the present research. In one of the few studies that used

psychologists exclusively as part of an industrial selection procedure (air pilots), candidates were rated on a 14 variable matrix on the basis of a clinical integration of paper-and-pencil tests. In what he describes as a "craftsman's job (p. 174)", Trankell (1959) describes how the integration of tests by a competent psychologist yields excellent results in terms of decision accuracy. He argues for the intelligent use of tests as predictors indicating that, rather than arguing relative merits, the strengths of each should be combined.

Summary: Literature Review

1. The general area of clinical judgement has been well researched specifically from the perspectives of predictive validity and model building. The area of clinical judgement in executive appraisal is only sparsely researched and the nature of that research has been primarily predictive validity studies of interviewing and assessment centers.

2. Clinical judgements, although they may be configural in nature, are adequately described by a linear model.

3. The linear model, whether it be used in a bootstrapping or traditional predictive manner, is at least the most accurate method of combining mathematically represented information for decision making. Even when beta weights are estimated or applied randomly, they better or equal a human judge working with the same information.

4. There is little research on the reliability of clinical

judgement. This is particularly true of convergence. What reliability studies that have been done have been concerned with concensus and/or stability. Convergence studies, when they have been attempted, have dealt with a similar data base (test or interview) or have been confounded with stability and/or concensus.

5. The majority of the research on clinical judgement, particularly that dealing with model building and predictive validity, would rate low in ecological (external) validity (Snow, 1974). If one views generalizability as a function of representativeness (Snow, 1974), the majority of the studies cited have been well off target. Typically, clinicians are required to rate subjects on variables that are foreign to them, using criteria and rating scales totally alien to their usual method, and are then critiqued for off-target behavior.

6. There exists at present no study which investigates the convergence of clinical judgement in a natural setting. This is particularly true of a natural, applied, vocational setting. Reliability is an extremely important, albeit ignored, concept in clinical judgement research (Goldberg & Werts, 1966). It should be noted that validity is unknown if the problems of reliability have not been accounted for. At present, the apple cart appears to have usurped the horse!

CHAPTER III

EXPERIMENTAL DESIGN

Clinician Sample

The three clinicians involved in this study are all professional staff of A. W. Fraser & Associates, a medium-sized, locally-owned industrial psychology and management consulting firm. Clinician #1, the chief psychologist, holds psychologist registration in three Canadian provinces, has over 12 years experience in executive appraisal and many more years of clinical experience. Clinician #2 has a B. A. (Hon) degree in psychology and over five years experience in executive appraisal. He was originally trained in executive appraisal techniques by Clinician #1 and was supervised very closely for the first three years in what might be described as an intensive and very highly supervised clinical-industrial internship. Clinician #3 is also a registered psychologist and has three years experience in industrial and executive appraisal. His most recent two years of experience have been obtained as a staff member of A. W. Fraser & Associates.

Subject Sample

Subjects utilized consist of recruitment and comprehensive appraisal candidates processed by the clinicians of A. W. Fraser & Associates from a time beginning with the inauguration of this study and ending when each clinician has rated at least twenty candidates. This covers the period March 1975-December 1975. Recruitment candidates are those candidates who have applied for

executive positions through the recruiting division of A. W. Fraser & Associates; comprehensive appraisal candidates are subjects sent to A. W. Fraser & Associates for assessment by their own companies in order to assess future development potential within that company.

Procedure

Definition of Traits

The definition of traits or characteristics of concern to the three clinicians of A. W. Fraser & Associates in assessing executive talent, were arrived at by a process of concensus by the three clinicians involved. Concensus was obtained on the number and name of the characteristics that 'make the difference' in executive performance and on the definition of these characteristics (Appendix 1). The three rating scales (Appendices 2, 3, & 4) used to quantify these characteristics had been in informal use in the organization previously but were modified to encompass the 18 key characteristics arrived at by concensus and the three information sources (test, interview, & test-interview).

Experimental Procedures

1. After completion of each assessment interview, the clinician completed the Interview Rating Form (Appendix 2) for the individual interviewed. This completed rating form was immediately returned to the office secretary for safekeeping and was not further available to the clinician.

2. The subject was administered the following tests as part of the appraisal battery: Differential Aptitude Test (Verbal and

Abstract); Wonderlic Personnel Test; Watson-Glaser Critical Thinking Appraisal; Test of Business Judgement; Test of Practical Judgement; Supervisory Practices Test; Management Aptitude Inventory; Holland Vocational Preference Inventory; Edwards Personal Preference Schedule; and the California Psychological Inventory (see Appendix 7 for summary description of tests). These tests comprise the usual executive assessment test battery utilized by the staff of A. W. Fraser & Associates; infrequently, additional tests are added to this battery.

3. The clinician was provided with a copy of the profiled results from all tests administered. Using the test results and interview impressions, the clinician completed the Interview + Test Rating Form (Appendix 3) for that candidate. This completed rating form was immediately returned to the office secretary for safe-keeping and was not further available to the clinician.

4. Approximately two months after the clinician had completed his required number of cases, he was provided with the test profiles from every subject he had previously rated. These profiles were made available to the clincian singly, in random order, and without identifying demographic information. The clinician then completed the Test Rating Form (Appendix 4) for each subject individually. This rating form was returned to the office secretary who collated the three rating forms from each subject.

Analysis Procedure

Ratings for each of the 18 characteristics variables (Appendix 1) for each of the three rater conditions (test, interview, and test+

interview) were analyzed by a one-way analysis of variance with repeated measures (ANOVA). This was done for each clinician individually and for all clinicians combined. If an F ratio obtained exceeded chance, individual comparisons between rater conditions were undertaken by the Newman-Keuls method of multiple comparisons. The reliability of the three ratings of each characteristic (factor) were also calculated as per a procedure outlined by Winer (1971, p. 290) and Ferguson (1971).

Experimental Hypotheses

There will be no significant differences between the means of the results obtained by any of the three assessment methods for any of the 18 characteristics for any of the three clinicians.

Limitations of the Study

This study is concerned with the convergence of clinical judgement across information sources with subject and rated characteristics held constant. Limitations then are limitations imposed by this restricted perspective.

1. No information will be available regarding the predictive validity of clinical judgement. This is not a study of predictive validity in clinical judgement, but rather a study of a specialized aspect of the process of clinical judgement.

2. Subjects were not randomly assigned to clinicians. Although no overt bias is present in subject assignment at A. W. Fraser & Associates, systematic covert bias in subject assignment cannot be excluded from consideration. In actual prac-

tice, each clinician is assigned to certain specific assignments based on his time availability and would see all subjects associated with that particular assignment. Snow (1974) would see this as the compromise that must occur between ecological validity on one hand and rigor of experimental design on the other.

3. Subject (client) selection was not random. Subjects can be considered to be representative of the types of clients who undertake executive appraisal.

4. All clinicians are male and all of the subjects are male. This may preclude generalizability of results to female populations.

5. Clinicians are not of equal training and experience. Although this has been seldom realized in a study of clinical judgement, there is a possible, but undetermined, effect on the generalizability of research findings. It is possible to investigate differences between clinicians but clinician sample size is far too small to investigate the effects of clinicians' characteristics on judgements of subject characteristics.

6. The possibility of clinicians' remembering profiles from the test + interview condition when they rated profiles in the test only condition is remote. It is, however, a possible weakness of design. The two month delay and the volume of work processed in that two month period did much to minimize this possibility.

CHAPTER IV

RESULTS

In this chapter, data pertaining to each of the clinicians by factor by rating condition interactions are presented. Results are organized by factor and are presented for each of the three clinicians in each of the three assessment conditions.

Definition of Terms

Since several terms will be used extensively in summarizing data analysis, a description of these terms, as they apply specifically to the present study, is given below:

F Ratio. Since the design utilized in this study involves a one-way analysis of variance with repeated measures, the ratio: F = Mean Square Treatment/Mean Square Residual is appropriate (Winer, 1971, p. 267).

Significant. Alpha is equal to .05.

Reliability (R). The reliability coefficient (R) is a simple proportion which represents the proportion of obtained variance that is true variance. For example, if R = .80, it means that 80% of the variation in the measurements is due to variation in the true score (real differences) with the remaining 20% variation due to error (Ferguson, 1971).

Unadjusted Reliability - Single Source (R1). The reliability of one estimate by one clinician of a single factor.

Unadjusted Reliability - Pooled Source (Rk). The reliability of the mean of the pooled or combined estimates of a single factor

by one clinician. This is frequently referred to as the Spearman-Brown reliability measure (Winer, 1971, p. 286).

Adjusted Reliability - Single Source (R*1). The reliability of one estimate by one clinician of a single factor after removal of mean differences between rating conditions as a source of error (Winer, 1971, p. 290).

Adjusted Reliability - Pooled Source (R*k). The reliability of the mean of the pooled or combined estimates of a single factor by one clinician after removal of mean differences between rating conditions as a source of error (Winer, 1971, p. 290).

The adjusted reliability coefficients R*1 and R*k are concerned with pegging or anchoring of the mid-points that a judge or rater appears to be using in estimating performance or ability on any given factor or trait. For example, if judges grading ten examination papers maintain essentially the same rank order so far as their grades are concerned, but differ in the actual values they assign, the use of an adjusted reliability estimate just described may be appropriate. The reliability model which removes mean differences is used when both means and variances are an important interpretation consideration from the perspective of error sources.

In discussion of reliability in this chapter, the adjusted reliability (R*1 and R*k) will be used predominantly, although both adjusted and unadjusted reliability estimates are presented in table form for reference. For purposes of the discussion of convergence, each of the reliability estimates just described

(R1, Rk, R*1, and R*k) can be considered as important and will be presented within the context of interpretation for each factor individually.

The relationship between R1 and Rk or R*1 and R*k may be expressed as: Rk = 3R1 (1 + 2R1) or R*k = 3R*1 (1 + 2R*1). This means that as R1 or R*1 approach one as an absolute value, R1 approaches Rk and R*1 approaches R*k.

Factor 1: Intelligence

Factor 1 has been defined as "the basic ability to learn and understand" (Appendix 1). In this study, aspects of this factor are sampled by clinical interpretation of psychometric tests such as the Wonderlic Personnel Test, Watson-Glaser Critical Thinking Appraisal, and the Differential Aptitude Tests (Abstract and Verbal) as well as by interview expertise.[1]

Table 1 presents the one-way analysis of variance with repeated measures (ANOVA) performed between the results obtained from each of the three assessment conditions for each of the clinicians individually. As is evident from Table 1, there is a significant degree of parallelism between the results obtained in each assessment category. This is true for all three clinicians. None of the *F* ratios obtained are sufficiently large to warrant further between-groups comparisons.

Tables 2 and 3 summarize the reliabilities, means, standard deviations associated with Factor 1. As would be expected on the basis of the previously mentioned *F* test, there is a marked similarity in both the means and standard deviations of the scores in each of the three assessment conditions for all three clinicians.

Clinicians differ markedly in the reliability of their decisions made with respect to levels of intelligence. Clinicians #1

---

[1]For each of the 18 factors, the tests which are indicated as being clinically combined for purposes of measuring these factors are as indicated by the three clinicians.

and #3 obtain a single measure reliability of approximately .50 with a pooled source Rk greater than .70. The single source R for Clinician #2 is so low as to cause concern for purposes of prediction. Even when one pools estimates (R*k), a value of only .32 is obtained, lower even than the R*1 for either of the other two clinicians. If this R value is in fact typical for all occasions, one should anticipate a low predictive validity of intelligence ratings made across information sources for Clinician #2. One might expect predictably unpredictable predictions!

TABLE 1

One Way Analysis of Variance with Repeated Measures

Factor 1: Intelligence

Rater	Source of Variation	Sums of Squares	df	F	p	p*
1	Between	20.18	19			
	Within	12.00	40			
	Treatment	1.23	2	2.18	.13	.16
	Residual	10.77	38			
	Total	32.18	59			
2	Between	9.99	23			
	Within	15.33	48			
	Treatment	1.78	2	3.02	.06	.09
	Residual	13.56	46			
	Total	25.32	71			
3	Between	29.43	29			
	Within	18.67	60			
	Treatment	1.27	2	2.11	.13	.16
	Residual	17.40	58			
	Total	48.10	89			

p* = Conservative probability of F which makes allowances for unequal covariances among correlated measures.

TABLE 2

Unadjusted and Adjusted Reliability Estimates

Factor 1: Intelligence

Clinician	Source	Unadjusted Reliability	Adjusted Reliability
1	Single	.46 (R1)	.48 (R*1)
1	Pooled	.72 (Rk)	.73 (R*k)
2	Single	.11 (R1)	.14 (R*1)
2	Pooled	.26 (Rk)	.32 (R*k)
3	Single	.43 (R1)	.44 (R*1)
3	Pooled	.69 (Rk)	.70 (R*k)

TABLE 3

Means and Standard Deviations

Factor 1: Intelligence

Clinician	Rating Condition	Mean	Standard Deviation
1	Interview	4.40	.58
1	Test	4.20	.81
1	Combined	4.55	.74
2	Interview	4.71	.54
2	Test	4.37	.70
2	Combined	4.71	.45
3	Interview	4.50	.50
3	Test	4.53	.76
3	Combined	4.27	.85

Factor 2: Common Sense

Factor 2 is described as "the degree of ability to reach quick, practically effective decisions about uncomplicated situations where sound judgement depends primarily on accumulated life and work experience, established precedent and procedures, etc." (Appendix 1). In this study, "common sense" is sampled by the clinical interpretation of tests such as Management Aptitude Inventory, California Psychological Inventory, and The Test of Practical Judgement in addition to interview evaluation.

Table 4 summarizes the ANOVA pertaining to Factor 2 for each of the three clinicians. For Clinicians #1 and #3, the differences in the diagnoses made between information sources are not significant. For Clinician #2 the differences in the diagnosis made between information sources are significant ($F = 4.48$, $p = .02$) and individual comparisons between groups are warranted. A Newman-Keuls multiple comparison between the three means (Winer, 1971, p. 217) indicates that the mean of the interview group is significantly greater than the mean of the test group and that the mean of the combined group is also significantly greater than the mean of the test group. There is no significant difference between the means of interview and combined groups for Clinician #2. It appears that subjects rated by Clinician #2 were rated significantly lower in the test condition than in either of the other two assessment conditions.

From Table 5, we see that these mean differences between

groups for Clinician #2, although significant, are not great, being in the order of .5. It is noteworthy that the standard deviation of the test condition for Clinician #2 is greater than that observed in either of the other two assessment conditions. The standard deviation of the test condition most closely parallels that of the combined assessment condition where one might expect test results to exert a moderating influence on the interview impressions.

Reliability values associated with Factor 2 for the three clinicians are moderate with R*1's in the order of .40 and R*k's in the order of .68. By more than tripling the amount of time required for purposes of evaluation, variance error is reduced by approximately 30%. A subject is appraised slightly differently in "common sense" depending on the assessment condition in which he is viewed. Particularly with Clinician #2, a candidate might be downrated somewhat if seen only in the test assessment condition.

TABLE 4

One Way Analysis of Variance with Repeated Measures

Factor 2: Common Sense

Rater	Source of Variation	Sums of Squares	df	F	p	p*
1	Between	12.85	19			
	Within	9.33	40			
	Treatment	.43	2	.93	.41	.35
	Residual	8.90	38			
	Total	22.18	59			
2	Between	25.11	23			
	Within	22.67	48			
	Treatment	3.69	2	4.48	.02	.05
	Residual	18.97	46			
	Total	47.78	71			
3	Between	47.39	29			
	Within	28.00	60			
	Treatment	2.49	2	2.83	.07	.10
	Residual	25.51	58			
	Total	75.39	89			

p* = Conservative probability of F which makes allowances for unequal covariances among correlated measures.

TABLE 5

Unadjusted and Adjusted Reliability Estimates

Factor 2: Common Sense

Clinician	Source	Unadjusted Reliability	Adjusted Reliability
1	Single	.39 (R1)	.39 (R*1)
1	Pooled	.65 (Rk)	.65 (R*k)
2	Single	.30 (R1)	.35 (R*1)
2	Pooled	.57 (Rk)	.62 (R*k)
3	Single	.45 (R1)	.48 (R*1)
3	Pooled	.71 (Rk)	.73 (R*k)

TABLE 6

Means and Standard Deviations

Factor 2: Common Sense

Clinician	Rating Condition	Mean	Standard Deviation
1	Interview	3.85	.65
1	Test	3.80	.60
1	Combined	4.00	.55
2	Interview	3.62	.56
2	Test	3.12	.88
2	Combined	3.58	.86
3	Interview	3.83	.73
3	Test	3.57	.99
3	Combined	3.43	.95

Factor 3: Oral Communication

Factor 3 is described by the clinicians involved in the study as "the degree of clarity and ease with which an individual expresses himself in face-to-face discussion" (Appendix 1). In this study, aspects of interpersonal effectiveness are sampled by interpretation of the California Psychological Inventory: Section I, and by interview evaluation.[1]

As evidenced by Table 7, the F ratios obtained for each of the three clinicians were not significant. Variances within groups and between groups were essentially the same. From Table 9, we see that this similarity is further evidenced by the close similarity of means and variances within each clinician cluster.

Reliability coefficients R*1 and R*k are not high, particularly for Clinicians #2 and #3. Although mean differences between rating conditions appear to cancel each other out as evidenced by the low F Ratios obtained, the effect of differential rankings on the R*1 and R*k values is considerable. Particularly for Clinicians #2 and #3, the reliability of any single estimate of oral communication ability (R*1) is so low as to have a great deal more of the prediction accountable for by error than is accountable for by true variation.

It is noteworthy that, although Clinician #3 indicated that he could not rate oral communication in the test condition, the other two clinicians were able to do so with results comparable to

---

[1]Clinician #3 did not rate Factor #3 in the test condition. He indicated that this was not normal procedure for him.

their ratings in the other two assessment conditions. However, it does not appear that test information regarding oral communication exerts much of a moderating influence vis-à-vis the distinctions between interview and combined scores for any clinician; they are highly parallel.

TABLE 7

One Way Analysis of Variance with Repeated Measures

Factor 3: Oral Communication

Rater	Source of Variation	Sums of Squares	df	F	p	p*
1	Between	16.73	19			
	Within	12.00	40			
	Treatment	.03	2	.05	.95	.82
	Residual	11.97	38			
	Total	28.73	59			
2	Between	22.17	23			
	Within	29.33	48			
	Treatment	.58	2	.47	.63	.50
	Residual	28.75	46			
	Total	51.50	71			
3	Between	20.60	29			
	Within	12.00	30			
	Treatment	.07	1	.16	.69	.69
	Residual	11.93	29			
	Total	32.60	59			

p* = Conservative probability of F which makes allowances for unequal covariances among correlated measures.

TABLE 8

Unadjusted and Adjusted Reliability Estimates

Factor 3: Oral Communication

Clinician	Source	Unadjusted Reliability	Adjusted Reliability
1	Single	.39 (R1)	.39 (R*1)
1	Pooled	.66 (Rk)	.64 (R*k)
2	Single	.16 (R1)	.15 (R*1)
2	Pooled	.37 (Rk)	.35 (R*k)
3	Single	.27 (R1)	.27 (R*1)
3	Pooled	.44 (Rk)	.42 (R*k)

TABLE 9

Means and Standard Deviations

Factor 3: Oral Communication

Clinician	Rating Condition	Mean	Standard Deviation
1	Interview	3.55	.74
1	Test	3.60	.66
1	Combined	3.55	.70
2	Interview	3.46	.86
2	Test	3.62	.90
2	Combined	3.67	.74
3	Interview	4.33	.91
3	Test	----	--
3	Combined	4.27	.51

Factor 4: Self-Starting Work Drive

Factor 4 is defined as "the degree to which an individual characteristically keeps himself continuously occupied in work related activities without need of stimulation from his supervisor" (Appendix 1). In this study, aspects of this factor are sampled by an interpretation of the Management Aptitude Inventory, Vocational Preference Inventory and California Psychological Inventory subscales, as well as by Interview evaluations.

Table 10 summarizes the ANOVA pertaining to Factor 4 for each of the three clinicians. As is evident, significant F ratios were obtained for Clinicians #1 and #3. In both cases, a Newman-Keuls multiple comparison between the respective mean differences, indicates that the mean of the interview group is significantly higher than the mean of the test group. For Clinicians #1 and #3, it seems that candidates impress as having more self-starting work drive when assessed by interview than when assessed by tests. There is also more variance in rating this factor in the test condition indicating that interview ratings are much more tightly clustered around the mean values (little inter-individual variation). R values are acceptably high with R*1 accounting for approximately 50% of the overall variance in all cases. R*k, which combines estimates from all rating conditions, improves on R*1 by approximately 20%. In practice, Factor 4 could probably be rated by any single method with acceptable results.

TABLE 10

One Way Analysis of Variance with Repeated Measures

Factor 4: Self-Starting Work Drive

Rater	Source of Variation	Sums of Squares	df	F	p	p*
1	Between	37.52	19			
	Within	23.33	40			
	Treatment	5.20	2	5.45	.008	.03
	Residual	18.13	38			
	Total	60.85	59			
2	Between	43.99	23			
	Within	21.33	48			
	Treatment	2.19	2	2.64	.08	.12
	Residual	19.14	46			
	Total	65.32	71			
3	Between	60.49	29			
	Within	46.00	60			
	Treatment	9.62	2	7.67	.001	.009
	Residual	36.38	58			
	Total	106.49	89			

p* = Conservative probability of F which makes allowances for unequal covariances among correlated measures.

TABLE 11

Unadjusted and Adjusted Reliability Estimates

Factor 4: Self-Starting Work Drive

Clinician	Source	Unadjusted Reliability	Adjusted Reliability
1	Single	.44 (R1)	.51 (R*1)
1	Pooled	.70 (Rk)	.76 (R*k)
2	Single	.52 (R1)	.55 (R*1)
2	Pooled	.77 (Rk)	.78 (R*k)
3	Single	.36 (R1)	.44 (R*1)
3	Pooled	.63 (Rk)	.70 (R*k)

TABLE 12

Means and Standard Deviations

Factor 4: Self-Starting Work Drive

Clinician	Rating Condition	Mean	Standard Deviation
1	Interview	3.95	.67
1	Test	3.25	1.18
1	Combined	3.45	.97
2	Interview	2.92	.64
2	Test	3.21	1.08
2	Combined	3.33	1.03
3	Interview	3.90	.70
3	Test	3.10	1.04
3	Combined	3.47	1.28

Factor 5: Interpersonal Effectiveness

Factor 5 is defined as "the level of effectiveness the individual demonstrates in day-to-day dealings with others with regard to gaining and maintaining their respect for his ideas and opinions, their confidence in his integrity, and their general feeling of good will" (Appendix 1). Aspects of this factor are appraised by the California Psychological Inventory, Vocational Preference Inventory, Edwards Personal Preference Schedule, Management Aptitude Inventory as well as by interview evaluations.

From Table 13, we see that significant F ratios were obtained only for Clinician #2. A Newman-Keuls comparison between mean differences indicates that the mean of the interview rating condition is significantly higher than both of the other two means. Subjects are rated significantly higher in interpersonal effectiveness during interview than when they are rated in either the test or combined condition. It seems likely that test information exerts a moderating influence on the interview evaluations when the combined rating is made. Combined ratings more closely parallel those of the test condition with respect to the pegging of mean values.

Although the results for Clinician #3 do not indicate a significant F, R values are very low. This indicates that, although deviations made over the total group within conditions appear to cancel one another out, ratings of individuals between conditions vary greatly. Even the R*k value of .36 is only at a level equal

to the R*1 value for the other two clinicians. More than three times the effort for Clinician #3 is required to match the reliability estimate for a single occasion for each of the other two clinicians. One should anticipate inconsistent predictions on interpersonal effectiveness for Clinician #3.

TABLE 13

One Way Analysis of Variance with Repeated Measures

Factor 5: Interpersonal Effectiveness

Rater	Source of Variation	Sums of Squares	df	F	p	p*
1	Between	16.67	19			
	Within	14.67	40			
	Treatment	.63	2	.86	.43	.37
	Residual	14.04	38			
	Total	31.33	59			
2	Between	22.54	23			
	Within	19.33	48			
	Treatment	4.08	2	6.16	.004	.02
	Residual	15.25	46			
	Total	41.88	71			
3	Between	21.16	29			
	Within	29.33	60			
	Treatment	2.16	2	2.30	.11	.14
	Residual	27.18	58			
	Total	50.49	89			

p* = Conservative probability of F which makes allowances for unequal covariances among correlated measures.

TABLE 14

Unadjusted and Adjusted Reliability Estimates

Factor 5: Interpersonal Effectiveness

Clinician	Source	Unadjusted Reliability	Adjusted Reliability
1	Single	.32 (R1)	.31 (R*1)
1	Pooled	.58 (Rk)	.58 (R*k)
2	Single	.32 (R1)	.39 (R*1)
2	Pooled	.59 (Rk)	.66 (R*k)
3	Single	.14 (R1)	.16 (R*1)
3	Pooled	.33 (Rk)	.36 (R*k)

TABLE 15

Means and Standard Deviations

Factor 5: Interpersonal Effectiveness

Clinician	Rating Condition	Mean	Standard Deviation
1	Interview	3.35	.79
1	Test	3.45	.80
1	Combined	3.20	.51
2	Interview	3.54	.76
2	Test	3.00	.76
2	Combined	3.08	.64
3	Interview	3.70	.78
3	Test	3.33	.74
3	Combined	3.43	.67

Factor 6: Leadership Force

Factor 6 is described as "the amount of influence and dominance the individual habitually exerts over groups and persons he encounters" (Appendix 1). Aspects of this factor are appraised by the California Psychological Inventory, Management Aptitude Inventory and by interview evaluations.

It is encouraging to view the results from the appraisal of leadership force under each of the three different rating conditions. Not only are the F ratios small, but reliability measures, in both the individual and pooled cases, are encouragingly high. Leadership force appears to be rated symetrically both between and within rating conditions. Further, there do not appear to be any inter-rater differences with respect to the ratings of leadership force. Means, standard deviations (Table 18), and reliabilities (Table 17) are highly convergent for all three clinicians.

TABLE 16

One Way Analysis of Variance with Repeated Measures

Factor 6: Leadership Force

Rater	Source of Variation	Sums of Squares	df	F	p	p*
1	Between	34.40	19			
	Within	15.33	40			
	Treatment	.63	2	.82	.44	.38
	Residual	14.70	38			
	Total	49.73	59			
2	Between	48.44	23			
	Within	23.33	48			
	Treatment	1.19	2	1.24	.30	.28
	Residual	22.14	46			
	Total	71.78	71			
3	Between	71.96	29			
	Within	36.67	60			
	Treatment	1.16	2	.94	.40	.34
	Residual	35.51	58			
	Total	108.62	89			

p* = Conservative probability of F which makes allowances for unequal covariances among correlated measures.

TABLE 17

Unadjusted and Adjusted Reliability Estimates

Factor 6: Leadership Force

Clinician	Source	Unadjusted Reliability	Adjusted Reliability
1	Single	.55 (R1)	.55 (R*1)
1	Pooled	.79 (Rk)	.79 (R*k)
2	Single	.53 (R1)	.53 (R*1)
2	Pooled	.77 (Rk)	.77 (R*k)
3	Single	.50 (R1)	.50 (R*1)
3	Pooled	.75 (Rk)	.75 (R*k)

TABLE 18

Means and Standard Deviations

Factor 6: Leadership Force

Clinician	Rating Condition	Mean	Standard Deviation
1	Interview	3.15	.79
1	Test	3.40	1.02
1	Combined	3.25	.89
2	Interview	2.83	.80
2	Test	2.87	1.01
2	Combined	3.12	1.13
3	Interview	3.20	1.01
3	Test	3.47	1.12
3	Combined	3.40	1.14

Factor 7: Self-Reliance

Self-reliance is "the degree to which the individual carries out assigned responsibilities without seeking direction, help, encouragement and/or reassurance from co-workers" (Appendix 1). In this study, elements of this factor are assessed by interpretation of the Edwards Personal Preference Schedule, California Psychological Inventory, Management Aptitude Inventory, and by interview evaluation.

Table 19 summarizes the ANOVA done with respect to Factor 7. As noted, significant differences between means were observed only for Clinician #2. A Newman-Keuls multiple comparison of mean differences reveals that the mean of the scores obtained from the interview condition is greater than the mean of the scores obtained in the test condition. Subjects were typically rated higher in self-reliance in the interview condition. Once again, for Clinician #2, test results appear to moderate interview impressions since the mean of the test condition is not significantly different from the mean of the interview condition.

Reliability measures for clinicians vary considerably for Factor 7. Both Clinicians #2 and #3 obtain R*k values which are less than the R*1 value obtained by Clinician #1. With R*1 equal to approximately .25 for Clinicians #2 and #3, one might expect a considerable difference in prediction dependent on rating condition.

TABLE 19

One Way Analysis of Variance with Repeated Measures

Factor 7: Self-Reliance

Rater	Source of Variation	Sums of Squares	df	F	p	p*
1	Between	35.73	19			
	Within	18.67	40			
	Treatment	.70	2	.74	.48	.40
	Residual	17.97	38			
	Total	54.40	59			
2	Between	22.61	23			
	Within	28.00	48			
	Treatment	4.19	2	4.05	.02	.05
	Residual	23.81	46			
	Total	50.61	71			
3	Between	45.15	29			
	Within	50.00	60			
	Treatment	4.69	2	3.00	.06	.09
	Residual	45.31	58			
	Total	95.16	89			

p* = Conservative probability of F which makes allowances for unequal covariances among correlated measures.

TABLE 20

Unadjusted and Adjusted Reliability Estimates

Factor 7: Self-Reliance

Clinician	Source	Unadjusted Reliability	Adjusted Reliability
1	Single	.50 (R1)	.50 (R*1)
1	Pooled	.75 (Rk)	.75 (R*k)
2	Single	.19 (R1)	.23 (R*1)
2	Pooled	.41 (Rk)	.47 (R*k)
3	Single	.22 (R1)	.25 (R*1)
3	Pooled	.46 (Rk)	.50 (R*k)

TABLE 21

Means and Standard Deviations

Factor 7: Self-Reliance

Clinician	Rating Condition	Mean	Standard Deviation
1	Interview	3.70	.46
1	Test	3.65	1.11
1	Combined	3.45	1.12
2	Interview	3.46	.81
2	Test	2.87	.66
2	Combined	3.08	.91
3	Interview	3.50	.81
3	Test	3.00	1.18
3	Combined	3.03	.98

Factor 8: Adaptability

Adaptability is defined as "the level of ability to cope comfortably with new and changing circumstances" (Appendix 1). In this study, aspects of this factor are appraised by tests such as the Edwards Personal Preference Schedule, the California Psychological Inventory, Vocational Preference Inventory, as well as by interview evaluations.

Table 22 summarizes the ANOVA relevant to Factor 8. As noted, no significant differences are evident, save for Clinician #3. A Newman-Keuls multiple comparison between mean differences for Clinician #3 indicates that the mean of the interview condition is significantly higher than the mean of the scores in either of the remaining two categories. In the same manner as was evident for Clinician #2 on Factors 7 and 5 and for Clinician #3 on Factor 4, the test protocols appear to exert a moderating influence on interview evaluations when a combined rating is undertaken.

The significant mean difference evidenced by Clinician #3 is combined with a low reliability ($R^*1 = .23$) indicating the very real possibility of differential diagnosis depending on the rating condition. For Clinician #1, although mean differences do not appear to be a large error source, considerable differences in ranking are apparent as reflected in the low value of $R^*1 = .30$ which is independent of the similarity or difference of mean pegging between groups. Clinician #2 obtained a R^*1 value which is considerably higher than even the R^*k value for the other two

clinicians. His single estimate of adaptability is encouragingly high and little is gained by combining all three methods.

TABLE 22

One Way Analysis of Variance with Repeated Measures

Factor 8: Adaptability

Rater	Source of Variation	Sums of Squares	df	F	p	p*
1	Between	17.52	19			
	Within	16.67	40			
	Treatment	1.23	2	1.52	.23	.23
	Residual	15.43	38			
	Total	34.18	59			
2	Between	47.11	23			
	Within	18.00	48			
	Treatment	1.86	2	2.65	.08	.12
	Residual	16.14	46			
	Total	65.11	71			
3	Between	41.29	29			
	Within	57.33	60			
	Treatment	13.75	2	9.15	.0003	.005
	Residual	43.58	58			
	Total	98.62	89			

p* = Conservative probability of F which makes allowances for unequal covariances among correlated measures.

TABLE 23

Unadjusted and Adjusted Reliability Estimates

Factor 8: Adaptability

Clinician	Source	Unadjusted Reliability	Adjusted Reliability
1	Single	.29(R1)	.30(R*1)
1	Pooled	.55(Rk)	.56(R*k)
2	Single	.60(R1)	.62(R*1)
2	Pooled	.82(Rk)	.83(R*k)
3	Single	.14(R1)	.23(R*1)
3	Pooled	.33(Rk)	.47(R*k)

TABLE 24

Means and Standard Deviations

Factor 8: Adaptability

Clinician	Rating Condition	Mean	Standard Deviation
1	Interview	3.30	.64
1	Test	3.10	.70
1	Combined	2.95	.86
2	Interview	3.33	.85
2	Test	3.04	1.06
2	Combined	2.96	.89
3	Interview	3.90	1.07
3	Test	3.00	.86
3	Combined	3.17	.97

Factor 9: Potential for Growth

Potential for growth is defined as "the degree of probability that an individual will develop the personal resources to cope with increasingly more complex and responsible work roles" (Appendix 1). In this study, potential for growth is appraised by a clinical integration of all information obtained by testing plus interview evaluations.

On evaluating the observations in Table 25 which summarizes the ANOVA for Factor 9, we see that a significant difference exists between the means of the three assessment conditions for Clinician #3. A Newman-Keuls multiple comparison of mean differences indicates that, as was the case for Clinician #3 on Factor 4, the mean of the interview assessment condition is significantly higher than the mean of the test assessment group. Once again, we see the moderating effect of test information on interview evaluations when rating in the combined condition. In the cases of Clinicians #1 and #2, a high degree of similarity is evident across rating conditions; no significant differences are evident.

Coupled with the significant differences in mean rating demonstrated by Clinician #3, we see a low R*1 associated with the estimation of Factor 9. Once again, the use of all three methods in obtaining an R*k = .65 for Clinician #3 only approximates the single source estimates obtained by Clinicians #1 and #2. Reliability estimates for Clinicians #1 and #2 are much more independent of assessment condition.

TABLE 25

One Way Analysis of Variance with Repeated Measures

Factor 9: Potential for Growth

Rater	Source of Variation	Sums of Squares	df	F	p	p*
1	Between	32.73	19			
	Within	12.00	40			
	Treatment	.63	2	1.06	.36	.32
	Residual	11.37	38			
	Total	44.73	59			
2	Between	48.65	23			
	Within	20.00	48			
	Treatment	1.03	2	1.25	.30	.28
	Residual	18.97	46			
	Total	68.65	71			
3	Between	38.99	29			
	Within	31.33	60			
	Treatment	3.90	2	4.11	.02	.05
	Residual	27.44	58			
	Total	70.32	89			

p* = Conservative probability of F which makes allowances for unequal covariances among correlated measures.

TABLE 26

Unadjusted and Adjusted Reliability Estimates

Factor 9: Potential for Growth

Clinician	Source	Unadjusted Reliability	Adjusted Reliability
1	Single	.61 (R1)	.61 (R*1)
1	Pooled	.83 (Rk)	.83 (R*k)
2	Single	.58 (R1)	.58 (R*1)
2	Pooled	.80 (Rk)	.80 (R*k)
3	Single	.34 (R1)	.38 (R*1)
3	Pooled	.61 (Rk)	.65 (R*k)

TABLE 27

Means and Standard Deviations

Factor 9: Potential for Growth

Clinician	Rating Condition	Mean	Standard Deviation
1	Interview	2.90	.62
1	Test	2.75	.99
1	Combined	2.65	.91
2	Interview	3.08	.86
2	Test	2.92	.95
2	Combined	3.21	1.08
3	Interview	3.57	.76
3	Test	3.07	.89
3	Combined	3.40	.92

Factor 10: Readiness to Learn

Readiness to learn is defined as "the individual's willingness to acquire new information, explore new ideas, methods, tasks, etc." (Appendix 1). In this study, it is appraised by tests such as the California Psychological Inventory, Vocational Preference Inventory, Wonderlic, and the Differential Aptitude Tests as well as by interview evaluations.

From an examination of Table 28, it appears that all clinicians experienced more difficulty in the rating of Factor 10 than they did with many of the other factors. Significant differences between rating conditions were evident for all three clinicians. A Newman-Keuls multiple comparison between means for each of the clinicians reveals considerable similarity in the differences exhibited. For Clinicians #1 and #2, the mean of the interview condition is significantly higher than the mean of the test condition. For Clinician #3, the mean of the interview condition is significantly greater than the mean of the test condition and the mean of the combined condition. Table 30 indicates that for all three clinicians, test results appear to be moderating interview impressions in the combined rating condition. For Clinician #3, this moderating effect is not great, resulting in the additional significant difference between interview and combined mean ratings.

Although significant F ratios were obtained for all clinicians, reliability estimates are not so uniform. Clinicians #1 and #2 parallel each other obtaining an R#1 value of approximately .47.

Clinician #3, as has been the case on Factors 8, 7, 5, and 3, obtains an R*1 value approximately one-half that of his counterparts. His degree of convergence between ratings is low.

TABLE 28

One Way Analysis of Variance with Repeated Measures

Factor 10: Readiness to Learn

Rater	Source of Variation	Sums of Squares	df	F	p	p*
1	Between	24.18	19			
	Within	16.67	40			
	Treatment	2.80	2	3.84	.03	.06
	Residual	13.87	38			
	Total	40.85	59			
2	Between	41.99	23			
	Within	26.00	48			
	Treatment	4.53	2	4.85	.01	.03
	Residual	21.47	46			
	Total	67.99	71			
3	Between	51.12	29			
	Within	59.33	60			
	Treatment	11.35	2	6.86	.002	.01
	Residual	47.98	58			
	Total	110.45	89			

p* = Conservative probability of F which makes allowances for unequal covariances among correlated measures.

TABLE 29

Unadjusted and Adjusted Reliability Estimates

Factor 10: Readiness to Learn

Clinician	Source	Unadjusted Reliability	Adjusted Reliability
1	Single	.41 (R1)	.45 (R*1)
1	Pooled	.67 (Rk)	.71 (R*k)
2	Single	.44 (R1)	.49 (R*1)
2	Pooled	.70 (Rk)	.74 (R*k)
3	Single	.21 (R1)	.27 (R*1)
3	Pooled	.44 (Rk)	.53 (R*k)

TABLE 30

Means and Standard Deviations

Factor 10: Readiness to Learn

Clinician	Rating Condition	Mean	Standard Deviation
1	Interview	3.35	.65
1	Test	2.85	.85
1	Combined	2.95	.86
2	Interview	3.83	1.07
2	Test	3.25	.92
2	Combined	3.37	.81
3	Interview	3.57	.99
3	Test	2.93	1.09
3	Combined	2.73	1.06

Factor 11: Management Level Planning and Problem Solving

Factor 11 is described as "the individual's ability to recognize the full depth and breadth of situations and problems and to consider the longer range, as well as the here-and-now consequences of their change or resolution" (Appendix 1). In this study, Factor 11 is appraised by the Watson-Glaser Critical Thinking Appraisal, Differential Aptitude Tests: Verbal and Abstract, California Psychological Inventory, Edwards Personal Preference Inventory, as well as by interview evaluation.

In comparison to the results for Factor 10 just presented, the results for Factor 11 are encouraging. From Table 31, we see that no mean differences, for any of the three clinicians, are significantly different from each other. There is a high degree of convergence within each clinician by rating condition cluster.

Reliability estimates for Factor 11 are also very respectable with values of R*1 approximating .55 for all clinicians. Apparently, both in terms of mean variation and intra-rating condition convergence, Factor 11 is regarded similarly by all three clinicians for all three ratings.

TABLE 31

One Way Analysis of Variance with Repeated Measures

Factor 11: Management Problem Solving

Rater	Source of Variation	Sums of Squares	df	F	p	p*
1	Between	58.27	19			
	Within	18.67	40			
	Treatment	.23	2	.24	.79	.63
	Residual	18.43	38			
	Total	76.93	59			
2	Between	85.33	23			
	Within	34.67	48			
	Treatment	.58	2	.39	.68	.54
	Residual	34.08	46			
	Total	120.00	71			
3	Between	92.90	29			
	Within	42.00	60			
	Treatment	.20	2	.14	.87	.71
	Residual	41.80	58			
	Total	134.90	89			

p* = Conservative probability of F which makes allowances for unequal covariances among correlated measures.

TABLE 32

Unadjusted and Adjusted Reliability Estimates

Factor 11: Management Problem Solving

Clinician	Source	Unadjusted Reliability	Adjusted Reliability
1	Single	.65 (R1)	.64 (R*1)
1	Pooled	.85 (Rk)	.84 (R*k)
2	Single	.58 (R1)	.57 (R*1)
2	Pooled	.81 (Rk)	.80 (R*k)
3	Single	.54 (R1)	.53 (R*1)
3	Pooled	.78 (Rk)	.77 (R*k)

TABLE 33

Means and Standard Deviations

Factor 11: Management Problem Solving

Clinician	Rating Condition	Mean	Standard Deviation
1	Interview	2.45	.80
1	Test	2.55	1.24
1	Combined	2.40	1.28
2	Interview	2.87	.88
2	Test	3.04	1.40
2	Combined	3.08	1.50
3	Interview	2.77	.67
3	Test	2.67	1.30
3	Combined	2.67	1.53

Factor 12: General Energy Level

General energy level is "the level of physical vigor and vitality the individual will demonstrate in his everyday conduct" (Appendix 1). This factor is sampled by the Edwards Personal Preference Schedule, California Psychological Inventory, Management Aptitude Inventory as well as by interview evaluations.

Except for Clinician #3, clinicians do not differ in their mean ratings between rating conditions for Factor 12. The significant difference between mean scores for Clinician #3, which is summarized in Table 34, is again indicative of a difference in mean pegging across rating conditions. A Newman-Keuls multiple comparison between the results of Clinician #3 indicates that, as was the case on many other factors, the mean of the interview rating condition is significantly higher than the mean of the test condition. Apparently, candidates are rated "more generously" in the interview condition than they are in the test condition.

Although Clinician #3 differs from his two counterparts in mean differences between rating conditions, he differs very little in obtained reliability estimates on Factor 12. All clinicians obtain R*1 values of approximately .28 indicating considerable ranking differences between rating conditions. With such a low R*1 value, differential diagnosis is a considerable possibility.

TABLE 34

One Way Analysis of Variance with Repeated Measures

Factor 12: General Energy Level

Rater	Source of Variation	Sums of Squares	df	F	p	p*
1	Between	10.18	19			
	Within	10.66	40			
	Treatment	.70	2	1.33	.27	.26
	Residual	9.97	38			
	Total	20.85	59			
2	Between	23.99	23			
	Within	21.33	48			
	Treatment	.53	2	.58	.56	.45
	Residual	20.80	46			
	Total	45.31	71			
3	Between	23.83	29			
	Within	24.67	60			
	Treatment	2.60	2	3.42	.04	.07
	Residual	22.07	58			
	Total	48.50	89			

p* = Conservative probability of F which makes allowances for unequal covariances among correlated measures.

TABLE 35

Unadjusted and Adjusted Reliability Estimates

Factor 12: General Energy Level

Clinician	Source	Unadjusted Reliability	Adjusted Reliability
1	Single	.25 (R1)	.26 (R*1)
1	Pooled	.50 (Rk)	.51 (R*k)
2	Single	.31 (R1)	.30 (R*1)
2	Pooled	.57 (Rk)	.57 (R*k)
3	Single	.25 (R1)	.28 (R*1)
3	Pooled	.50 (Rk)	.54 (R*k)

TABLE 36

Means and Standard Deviations

Factor 12: General Energy Level

Clinician	Rating Condition	Mean	Standard Deviation
1	Interview	3.65	.48
1	Test	3.40	.58
1	Combined	3.60	.66
2	Interview	3.58	.76
2	Test	3.50	.64
2	Combined	3.71	.93
3	Interview	4.07	.68
3	Test	3.67	.70
3	Combined	3.77	.76

Factor 13: Efficiency of Application

Efficiency of application is defined as "the economic and productive organization and application of work time and effort" (Appendix 1). It is sampled by the Management Aptitude Inventory, California Psychological Inventory, Vocational Preference Inventory, Test of Practical Judgement as well as by interview evaluations.

Table 37 summarizes the ANOVA associated with Factor 13 for all three clinicians. As is evident from the table, there are no significant differences within each clinician by rating condition cluster. Table 38 summarizes the intra-rater reliability estimates for Factor 13. For Clinicians #1 and #3, it appears that the absence of significant mean differences between rating conditions is complimented by substantial R*1 values approximating .50. Clinician #2, however, does not match this level of convergence obtaining an R*1 value of only .16. This value would make the reliability of any individual decision, based on any one rating condition, tenuous. As should be expected, R*k values are in close correspondence with those obtained for R*1. However, even the R*k value of .36 for Clinician #2 is of concern for prediction purposes.

TABLE 37

One Way Analysis of Variance with Repeated Measures

Factor 13: Efficiency of Application

Rater	Source of Variation	Sums of Squares	df	F	p	p*
1	Between	24.40	19			
	Within	14.00	40			
	Treatment	.40	2	.56	.58	.46
	Residual	13.60	38			
	Total	38.40	59			
2	Between	20.32	23			
	Within	26.67	48			
	Treatment	.86	2	.77	.47	.39
	Residual	25.80	46			
	Total	46.99	71			
3	Between	66.99	29			
	Within	28.67	60			
	Treatment	.69	2	.71	.49	.40
	Residual	27.98	58			
	Total	95.65	89			

p* = Conservative probability of F which makes allowances for unequal covariances among correlated measures.

TABLE 38

Unadjusted and Adjusted Reliability Estimates

Factor 13: Efficiency of Application

Clinician	Source	Unadjusted Reliability	Adjusted Reliability
1	Single	.47 (R1)	.46 (R*1)
1	Pooled	.73 (Rk)	.72 (R*k)
2	Single	.16 (R1)	.16 (R*1)
2	Pooled	.37 (Rk)	.36 (R*k)
3	Single	.56 (R1)	.56 (R*1)
3	Pooled	.79 (Rk)	.79 (R*k)

TABLE 39

Means and Standard Deviations

Factor 13: Efficiency of Application

Clinician	Rating Condition	Mean	Standard Deviation
1	Interview	3.70	.64
1	Test	3.50	.97
1	Combined	3.60	.73
2	Interview	3.17	.80
2	Test	2.92	.86
2	Combined	2.96	.73
3	Interview	3.20	.75
3	Test	3.00	1.18
3	Combined	3.17	1.10

Factor 14: Self-Confidence

Self-confidence is described as "the degree of basic security the individual feels in his own ability to deal adequately with most situations and people he encounters" (Appendix 1). This factor is sampled by the California Psychological Inventory, Edwards Personal Preference Schedule, Management Aptitude Inventory and by interview evaluations.

As summarized in Table 40, significant differences between rating condition means are evident for Clinicians #2 and #3. A Newman-Keuls multiple comparison between mean differences reveals that, for both clinicians, the mean of the interview rating condition is significantly higher than the mean of the interview rating condition. For both of these clinicians, test results moderate interview evaluations to yield a combined rating lower than the interview rating but higher than the test rating. This is not the case for Clinician #1 where, if anything might be said about the statistically insignificant differences, it should be that interview evaluations moderate higher test ratings.

Reliability estimates for all the three clinicians range from barely acceptable (R*1 = .34) to quite credible (R*1 = .55). Roughly 20% of the error variance vis-à-vis reliability is controlled by averaging the results from all three procedures rated independently (R*k).

TABLE 40

One Way Analysis of Variance with Repeated Measures

Factor 14: Self-Confidence

Rater	Source of Variation	Sums of Squares	df	F	p	p*
1	Between	22.98	19			
	Within	10.67	40			
	Treatment	.90	2	1.75	.19	.20
	Residual	9.77	38			
	Total	33.60	59			
2	Between	26.00	23			
	Within	24.00	48			
	Treatment	3.58	2	4.04	.02	.06
	Residual	20.42	46			
	Total	50.00	71			
3	Between	35.65	29			
	Within	25.33	60			
	Treatment	2.82	2	3.63	.03	.07
	Residual	22.51	58			
	Total	60.99	89			

p* = Conservative probability of F which makes allowances for unequal covariances among correlated measures.

TABLE 41

Unadjusted and Adjusted Reliability Estimates

Factor 14: Self-Confidence

Clinician	Source	Unadjusted Reliability	Adjusted Reliability
1	Single	.54 (R1)	.55 (R*1)
1	Pooled	.78 (Rk)	.79 (R*k)
2	Single	.30 (R1)	.34 (R*1)
2	Pooled	.56 (Rk)	.61 (R*k)
3	Single	.39 (R1)	.42 (R*1)
3	Pooled	.66 (Rk)	.68 (R*k)

TABLE 42

Means and Standard Deviations

Factor 14: Self-Confidence

Clinician	Rating Condition	Mean	Standard Deviation
1	Interview	3.70	.64
1	Test	4.00	.89
1	Combined	3.85	.65
2	Interview	3.75	.83
2	Test	3.21	.76
2	Combined	3.54	.81
3	Interview	4.43	.80
3	Test	4.00	.86
3	Combined	4.20	.75

Factor 15: Supervisory Effectiveness

Supervisory effectiveness refers to "the individual's habitual effectiveness in directing, co-ordinating, and controlling subordinates in standard work settings" (Appendix 1). This factor is appraised by the California Psychological Inventory, Edwards Personal Preference Schedule, Management Aptitude Inventory, Supervisory Practices Test, Test of Practical Judgement and interview evaluations.

As has been the case for many other factors, there is a high convergence of mean ratings within each clinician by rating condition cluster. Candidates appear to be rated on the same yardstick in each of the three rating conditions.

If, as noted above, candidates are being rated with the same yardstick in each rating condition, they are not measured identically in each case. Individual case (R*1) reliability estimates for Clinicians #1 and #2 are only low to moderate (R*1 = .30 or .40). Clinician #3, however, is remarkably consistent in his ratings of supervisory effectiveness between rating conditions (R*1 = .56). For him, the possibility of differing diagnosis as a function of assessment condition is reduced.

TABLE 43

One Way Analysis of Variance with Repeated Measures

Factor 15: Supervisory Effectiveness

Rater	Source of Variation	Sums of Squares	df	F	p	p*
1	Between	28.18	19			
	Within	20.67	40			
	Treatment	1.90	2	1.92	.16	.18
	Residual	18.77	38			
	Total	48.85	59			
2	Between	21.78	23			
	Within	20.67	48			
	Treatment	1.36	2	1.62	.21	.21
	Residual	19.30	46			
	Total	42.44	71			
3	Between	67.83	29			
	Within	28.67	60			
	Treatment	.07	2	.07	.93	.80
	Residual	28.60	58			
	Total	96.50	89			

p* = Conservative probability of F which makes allowances for unequal covariances among correlated measures.

TABLE 44

Unadjusted and Adjusted Reliability Estimates

Factor 15: Supervisory Effectiveness

Clinician	Source	Unadjusted Reliability	Adjusted Reliability
1	Single	.38 (R1)	.40 (R*1)
1	Pooled	.65 (Rk)	.67 (R*k)
2	Single	.29 (R1)	.30 (R*1)
2	Pooled	.55 (Rk)	.56 (R*k)
3	Single	.56 (R1)	.56 (R*1)
3	Pooled	.80 (Rk)	.79 (R*k)

TABLE 45

Means and Standard Deviations

Factor 15: Supervisory Effectiveness

Clinician	Rating Condition	Mean	Standard Deviation
1	Interview	3.30	.78
1	Test	3.65	1.01
1	Combined	3.70	.84
2	Interview	2.87	.66
2	Test	2.54	.76
2	Combined	2.75	.83
3	Interview	2.87	.72
3	Test	2.83	1.24
3	Combined	2.80	1.08

Factor 16: Autonomy

Autonomy is described as "the degree of the individual's need to make his own decisions, regulate his own behavior, be his own boss, etc." (Appendix 1). This factor is appraised by the Edwards Personal Preference Schedule, California Psychological Inventory and interview evaluations.

Table 46 indicates that there is a similarity in mean ratings between rating conditions for Clinicians #1 and #2. Clinician #3, however, obtains a very significant F ratio indicating differences between rating conditions with respect to mean ratings. A Newman-Keuls multiple comparison between the means of the three rating conditions indicates that the mean of the interview condition is significantly greater than the mean of the test and combined conditions. Test results appear to moderate interview evaluations very considerably when rating in the combined condition.

Although there are significant mean differences between rating conditions for Clinician #3, we see from Table 47 that, once mean differences are removed as a source of error, his convergence estimate (R#1) is considerably higher than that observed for the other two clinicians. This points up the necessity of considering both mean differences and intra-rater reliability when discussing convergence.

TABLE 46

One Way Analysis of Variance with Repeated Measures

Factor 16: Autonomy

Rater	Source of Variation	Sums of Squares	df	F	p	p*
1	Between	22.98	19			
	Within	20.00	40			
	Treatment	1.23	2	1.25	.30	.28
	Residual	18.76	38			
	Total	42,98	59			
2	Between	32.65	23			
	Within	28.67	48			
	Treatment	2.03	2	1.75	.19	.20
	Residual	26.64	46			
	Total	61.32	71			
3	Between	61.29	29			
	Within	40.67	60			
	Treatment	5.35	2	4.40	.02	.04
	Residual	35.31	58			
	Total	101.96	89			

p* = Conservative probability of F which makes allowances for unequal covariances among correlated measures.

TABLE 47

Unadjusted and Adjusted Reliability Estimates

Factor 16: Autonomy

Clinician	Source	Unadjusted Reliability	Adjusted Reliability
1	Single	.32 (R1)	.33 (R*1)
1	Pooled	.59 (Rk)	.59 (R*k)
2	Single	.31 (R1)	.33 (R*1)
2	Pooled	.58 (Rk)	.59 (R*k)
3	Single	.41 (R1)	.45 (R*1)
3	Pooled	.68 (Rk)	.71 (R*k)

TABLE 48

Means and Standard Deviations

Factor 16: Autonomy

Clinician	Rating Condition	Mean	Standard Deviation
1	Interview	3.35	.57
1	Test	3.20	.98
1	Combined	3.00	.89
2	Interview	3.08	.81
2	Test	2.71	.93
2	Combined	2.78	.97
3	Interview	3.37	1.05
3	Test	2.83	1.00
3	Combined	2.87	1.06

Factor 17: Responsibility

Factor 17 refers to "the degree to which the individual lives up to personal, professional, and business obligations he has tacitly or otherwise accepted" (Appendix 1). This is assessed by an interpretation of the California Psychological Inventory, Management Aptitude Inventory, Edwards Personal Preference Schedule and interview evaluations.

Table 49 summarizes the F tests associated with Factor 17. The results of Clinician #1 indicate a marginally significant difference between the means of the three assessment conditions. However, for Clinician #1, a Newman-Keuls multiple comparison between mean differences indicates that, although the overall F ratio is significant, no individual difference between mean pairs is great enough to be considered significant. Clinician #3 also obtains a significant F indicating significant overall differences between groups. Further, a Newman-Keuls multiple comparison reveals that the mean of the interview condition is greater than the mean of the combined rating condition. On referring to Table 51, it is surprising to note that the mean of the combined rating condition is lower than either of the test or interview rating conditions.

Intra-rater reliability (R*k) estimates are also moderate for all three clinicians for Factor 18 being in the order of .30 to .40. Apparently, candidates are rated differently in the three rating conditions, but differences in rating made in a positive direction are nearly equalled by differences in rating made in the negative

direction (low F and low R*1).

TABLE 49

One Way Analysis of Variance with Repeated Measures

Factor 17: Responsibility

Rater	Source of Variation	Sums of Squares	df	F	p	p*
1	Between	16.27	19			
	Within	16.67	40			
	Treatment	2.43	2	3.25	.05	.09
	Residual	14.23	38			
	Total	32.94	59			
2	Between	19.54	23			
	Within	13.33	48			
	Treatment	.58	2	1.05	.36	.31
	Residual	12.75	46			
	Total	32.87	71			
3	Between	35.12	29			
	Within	37.33	60			
	Treatment	6.75	2	6.41	.003	.02
	Residual	30.58	58			
	Total	72.45	89			

p* = Conservative probability of F which makes allowances for unequal covariances among correlated measures.

TABLE 50

Unadjusted and Adjusted Reliability Estimates

Factor 17: Responsibility

Clinician	Source	Unadjusted Reliability	Adjusted Reliability
1	Single	.26 (R1)	.30 (R*1)
1	Pooled	.51 (Rk)	.56 (R*k)
2	Single	.41 (R1)	.41 (R*1)
2	Pooled	.67 (Rk)	.67 (R*k)
3	Single	.24 (R1)	.30 (R*1)
3	Pooled	.49 (Rk)	.56 (R*k)

TABLE 51

Means and Standard Deviations

Factor 17: Responsibility

Clinician	Rating Condition	Mean	Standard Deviation
1	Interview	3.75	.62
1	Test	3.35	.73
1	Combined	3.30	.78
2	Interview	3.37	.63
2	Test	3.17	.69
2	Combined	3.33	.69
3	Interview	3.43	1.02
3	Test	3.03	.66
3	Combined	2.77	.84

Factor 18: General Suitability

Factor 18 is described as a "self explanatory" rating (Appendix 1) in that it refers to the overall suitability or overall rating of a candidate. It could be likened to the measure of general intelligence in intellectual assessments in that a composite is presumed.

Statistics associated with Factor 18 are somewhat alarming. Only Clinician #3 obtains a significant F and a further Newman-Keuls multiple comparison indicates that the mean of the interview assessment condition is significantly greater than the mean of either of the other two rating conditions.

It is the low R*1 values which disclose the most about the rating of this factor. Values range from a low of .05 to a high of .33, the lowest seen for any factor. This indicates considerable intra-individual ranking differences. Candidates are not viewed as uniform with respect to their overall suitability across rating conditions.

TABLE 52

One Way Analysis of Variance with Repeated Measures

Factor 18: General Suitability

Rater	Source of Variation	Sums of Squares	df	F	p	p*
1	Between	14.31	19			
	Within	26.67	40			
	Treatment	2.13	2	1.65	.20	.21
	Residual	24.53	38			
	Total	40,98	59			
2	Between	24.61	23			
	Within	30.00	48			
	Treatment	3.36	2	2.90	.06	.10
	Residual	26.64	46			
	Total	54.61	71			
3	Between	27.65	29			
	Within	26.00	60			
	Treatment	3.49	2	4.49	.01	.04
	Residual	22.51	58			
	Total	53.65	89			

p* = Conservative probability of F which makes allowances for unequal covariances among correlated measures.

TABLE 53

Unadjusted and Adjusted Reliability Estimates

Factor 18: General Suitability

Clinician	Source	Unadjusted Reliability	Adjusted Reliability
1	Single	.04 (R1)	.05 (R*1)
1	Pooled	.12 (Rk)	.14 (R*k)
2	Single	.19 (R1)	.22 (R*1)
2	Pooled	.42 (Rk)	.46 (R*k)
3	Single	.29 (R1)	.33 (R*1)
3	Pooled	.55 (Rk)	.59 (R*k)

TABLE 54

Means and Standard Deviations

Factor 18: General Suitability

Clinician	Rating Condition	Mean	Standard Deviation
1	Interview	3.45	.74
1	Test	3.45	1.02
1	Combined	3.05	.59
2	Interview	3.17	.62
2	Test	2.71	.89
2	Combined	2.71	.98
3	Interview	3.40	.66
3	Test	3.00	.82
3	Combined	2.97	.75

Inter-Rater Reliability: Test Condition

As an added precaution against the possibility of clinicians remembering test profiles already used in the combined rating condition when they were rating in the test condition, all clinicians rated all 74 test profiles (their own plus those of the other two clinicians). As noted in Chapter 3, each profile was rated individually, in random sequential order, without identifying demographic information. A side effect of using this blind rating approach is that it is possible to see how closely the three clinicians involved in this study rated the same profiles; i.e., it is possible to obtain a measure of the inter-rater reliability (concensus) of test condition assessment decisions as well as the intra-rater reliability estimates already presented. Inter-rater reliability is important because it can give us some idea of the consistency of three clinicians in rating similar profiles under similar situations. If inter-rater reliability is low, the problem of good predictions is further complicated. Not only would differences in rating situations be important in so far as the actual rating is concerned, but the rating made would also be extremely clinician-dependent. Although in this study, and in most real life assessment situations, a candidate is usually rated by only one clinician, it is interesting to note how much of the rating given is "clinician-dependent" and how much is "clinician-independent" with respect to assigned value. This says nothing about validity however, since high consistency does not necessarily lead to

prediction accuracy.

The inter-rater reliability indices of the three clinicians ratings done in the test rating condition for all 74 subjects on 17 factors are summarized in Table 55. Factor 3 is not presented since it will be recalled that Clinician #3 did not rate oral communication in the test condition mode.

It is seen that the R*1 inter-rater reliability estimates vary from a low of .19 to a high of .89 with a mean value of .62. With the exception of the R*1 value of .19 for Factor 12, all reliability estimates ≥ .50. Factor 12, as was evident from Tables 34-36, was a factor with which all clinicians had difficulty in cross-group ratings; intra-rater reliability indices were also very low. It seems that, even with a single category of information, clinicians differ in their interpretation of "general energy level" and/or how it is measured via psychometric profiles.

TABLE 55

Unadjusted and Adjusted Inter-Rater Reliability Estimates

Test Rating Condition

Factor	Unadjusted Reliability (R1)	Adjusted Reliability (R*1)
1	.61	.66
2	.77	.80
4	.87	.89
5	.48	.49
6	.67	.69
7	.66	.69
8	.65	.66
9	.53	.56
10	.55	.55
11	.87	.87
12	.19	.19
13	.53	.53
14	.42	.50
15	.50	.53
16	.74	.78
17	.56	.56
18	.54	.58

Factor Analysis of Test Condition Ratings

A factor analysis of each clinician's ratings on the 18 factors for all candidates (N = 74) rated in the test condition was undertaken. This procedure was deemed useful to assist in a discussion of the results just presented. It was thought useful to examine clusters of similar factor ratings made between candidates to establish possible communalities of ratings. It seems likely that, although factors have been presented as semantically and constructually idiosyncratic (Appendix 1), there are common ratings made on an individual between factors, i.e., ratings may be mutually interdependent.

With this in mind, a principal axis factoring with varimax orthogonal rotation was attempted with the results obtained from each clinician's rating of the 74 candidates on the 18 factors in the test condition. Since this factor analysis is not central to this results section, findings are detailed in Appendix 6 and are presented here in summary form only.

Using a criterion eigen value = 1.00, each clinician's original ratings based on 18 factors were found to load significantly on five major factors. The percentage of total variance of the original 18 factors accounted for by the five new factors ranges from 71% for Clinician #1 to 60% for Clinician #3.

The results of the factor analysis are presented individually by clinician. A tentative descriptive title for each of the five prime factors for each of the three clinicians is typed in brackets

immediately following the factors which appear to load significantly on that factor.

Factor Loadings

Clinician #1

FACTOR I: General intelligence + adaptability + potential for growth + readiness to learn + management level planning and problem solving. Total variance accounted for = 22% (INTELLECTUAL POTENTIAL).

FACTOR II: Oral communication + leadership force + interpersonal effectiveness + self-confidence + supervisory effectiveness. Variance accounted for = 18%. (INTERPERSONAL FORCEFULNESS).

FACTOR III: Efficiency of application + responsibility + general suitability. Variance accounted for = 17%. (RESPONSIBLE EFFICIENT WORK STYLE).

FACTOR IV: Self-starting work drive + general energy level. Variance accounted for = 9%. (WORK DRIVE).

FACTOR V: Common sense + self-reliance + autonomy. Variance accounted for = 9% (RESOURCEFULNESS).

Total variance accounted for by factors I - V = 71%.

Clinician #2

FACTOR I: General intelligence + oral communication + potential for growth + readiness to learn + management level planning and problem solving + adaptability. Variance accounted for = 22%. (INTELLECTUAL POTENTIAL).

FACTOR II: General energy level + responsibility + general suitability. Variance accounted for = 15%. (DIRECTED ENERGY).

FACTOR III: Self-reliance + self-confidence + autonomy. Variance accounted for = 13%. (RESOURCEFULNESS).

FACTOR IV: Common sense + interpersonal effectiveness + supervisory effectiveness. Variance accounted for = 10%. (INTERPERSONAL FORCEFULNESS).

FACTOR V: Self-starting work drive + efficiency of application. Variance accounted for = 8%. (GOAL DIRECTED WORK DRIVE).

Total variance accounted for by factors I - V = 68%.

Clinician #3

FACTOR I: Leadership force + general energy level + self-confidence + supervisory effectiveness. Variance accounted for = 14%. (DYNAMIC LEADERSHIP).

FACTOR II: Adaptability + potential for growth + readiness to learn + management level planning and problem solving. Variance accounted for = 14%. (POTENTIAL ABILITY).

FACTOR III: Self-reliance + efficiency of application + responsibility + general suitability. Variance accounted for = 13%. (RESOURCEFULNESS).

FACTOR IV: General intelligence + common sense. Variance accounted for = 9%. (PRACTICAL PROBLEM SOLVING).

FACTOR V: Self-starting work drive + autonomy. Variance accounted for = 9%. (INDEPENDENT WORK STYLE).

Total variance accounted for = 60%.

Summary

Clinician #1

Factors with significant differences between means: 4, 10, 17

Mean R*1 value for all 18 factors = .41; standard deviation = .14

Factors with R*1 = 0 - .30: 8, 12, 17, 18

Factors with R*1 = .31 - .60: 1, 2, 3, 4, 5, 6, 7, 10, 13, 14, 15, 16

Factors with R*1 = .61 - 1.00: 9, 11

Clinician #2

Factors with significant differences between means: 2, 5, 7, 10, 14

Mean R*1 value for all 18 factors =.37; standard deviation = .15

Factors with R*1 = 0 - .30: 1, 3, 7, 12, 13, 15, 18

Factors with R*1 = .31 - .60: 2, 4, 5, 6, 9, 10, 11, 14, 16, 17

Factors with R*1 = .61 - 1.00: 8

Clinician #3

Factors with significant differences between means: 4, 8, 9, 10, 11, 12, 14, 16, 17, 18

Mean R*1 value for 17 factors (excepting #3) = .38; standard deviation = .12

Factors with R*1 = 0 - .31: 3, 5, 7, 8, 10, 12, 17

Factors with R*1 = .31 - .60: 1, 2, 4, 6, 9, 11, 13, 14, 15, 16, 18

CHAPTER V

DISCUSSION

In this chapter, results pertaining to each of the clinician by factor by rating condition interactions will be discussed. Common themes will be examined by clinician and factor, an attempt will be made to explain significant results, the utility of convergence as a psychological construct will be examined, and suggestions for further research will be detailed.

Within the context of this study, convergence is probably best viewed as a condition affected by both mean differences and reliability estimates. It is possible to err with the ratings made, both from the point of view of the actual rating assigned to a candidate within any rating condition, and from the perspective of differences in rating of a candidate made across conditions. The first difference, which is often described as mean pegging error, can be considered to be a constant. Errors of this type would result in comparison errors when inflated scores from one group are compared to deflated scores from another. This type of error is unlikely to result in errors when considering an individual within any group since rankings are not changed (each person is being measured with the same, albeit incorrect, yardstick). Mean pegging errors are also very easy to correct since changing all raw rating scores from all groups to standard scores will standardize between groups.

From a consideration of the tables in Chapter 4, it is evident

that, even when the difference between rating condition means is very statistically significant, (e.g., Clinician #3 on Factor 8), the actual numerical differences between the significantly different mean-pairs is not great. Thus, from observing the tables in Chapter 4 once again, we see most of the raw score differences between mean-pairs that are significantly different are in the order of .50 - .80. Since actual ratings made on candidates are in whole numbers within the range 1 - 5, it is unlikely that differences across rating conditions for any candidate would exceed one. Thus, significant differences between the means of the ratings made in each category may not reflect practical differences between ratings from the perspective of actual judgements made about that candidate. What will be said differently about a candidate who scores 4 on a factor versus that said about a candidate who scores 5?

Low R*1 estimates are a reflection of low concurrence in subject ratings across rating conditions once mean error has been removed. Low R*1 values should be viewed more seriously than high F values since they cannot be eliminated by anything as simple as a standard score transformation.

Low R*1 values may be thought of as reflecting either or both of two possibilities: (1) basic clinician decision error of the type noted in the equation, TRUE SCORE = OBTAINED SCORE + ERROR or, (2) real differences inherent in the information available about a candidate as a result of sampling in either of the three appraisal conditions which would cause even a totally accurate clinician to

diagnose differentially dependent on assessment condition (i.e., real differences in the quantitative information available about a candidate). The first possibility is that referred to by researchers such as Little & Scheidman (1959) or Goldberg & Werts (1966). The second has been ignored in the literature.

It is this second possibility that is most frustrating for the researcher - and so face-saving for the clinician! It may be that differences in intra-rater reliability are differences, not due only to clinician error, but in differences in the ability assessed in each condition. This could also be thought of as a construct difference between factors which bear the same name in each of the three conditions. It may never be possible to separate these two types of "error", but it is wise to keep them in mind, particularly when discussing intra-rater reliability.

It seems logical to presume that, when all three clinicians obtain high R*1 values on the same factor, both types of error would be minimized. Similarly, when one or two clinicians obtain a high R*1 value on any factor, it is tenable to assume that the lower R*1 value of the other clinician(s) on the same factor reflects judgement errors (type 1) rather than real differences in the level of ability assessed by different methods (type 2). One would assume that the second type of error would be a constant between and within any given clinician by factor cluster with R*1 scores which are lower than the highest R*1 value obtained by any of the three clinicians being due to clinician decision error.

It is also logical to assume that, when R*1 values are low for all clinicians, or when there is considerable variability between the R*1 values of each clinician on the same factor, that both types of error are greater although the relationship between the magnitudes of the two types of error is indeterminate. It should be noted that inter-rater reliability estimates (Table 55) include only error of the first type (judgement error), since the same sources of information were available to all clinicians. This would be the essential difference in the interpretation of intra- versus inter-rater reliability estimates. With these different types of error in mind, let us examine intra- and inter-rater reliability in the present study.

Let us assume for the present that an acceptable level of intra-rater reliability would be approximately .50. In actual fact, the choice of any criterion value is always arbitrary representing a compromise between practical limitations and statistical desirability. With an R*1 value equal to approximately .50, we would assume that roughly 50% of the variance of any single estimate of any factor represented true variance with the remaining 50% being due to error of various types. Although the choice of .50 as a criterion value may appear somewhat lenient, it is realistic given the differences between statistical and practical significance vis-à-vis score assignment differences previously discussed.

If one examines each clinician's R*1 estimates across all

18 factors, we see that there are 12 factors where at least one clinician obtains an R*1 approximately equal to .50. These factors are Factor 4 (self-starting work drive), Factor 6 (leadership force), Factor 7 (self-reliance), Factor 8 (adaptability), Factor 9 (potential for growth), Factor 11 (management level planning and problem solving ability), Factor 13 (efficiency of application), Factor 14 (self-confidence), Factor 15 (supervisory effectiveness), Factor 1 (intelligence), Factor 2 (common sense), and Factor 10 (readiness to learn). In several cases, two or even three clinicians obtain these criterion R*1 values for the factor noted. Factors where no clinician achieves a criterion R*1 value are Factor 3 (oral communication), Factor 5 (interpersonal effectiveness), Factor 12 (general energy level), Factor 16 (autonomy), Factor 17 (responsibility), and Factor 18 (general suitability). For these factors, both types of error would be considerable.

As noted earlier, it is tenable to consider that, for the factors where one or more clinicians obtains an R*1 value approximating .50, the difference between this value and the R*1 value obtained by the other clinician(s) on the same factor is comprised primarily of clinician judgement error (type 1) rather than essential differences in the levels of ability measured (type 2). Each clinician is availed the same types of information about each of the candidates to be appraised as is every other clinician. Therefore, errors of the second type would be presumed to be a constant for all clinicians; possibly large, but still a constant.

If one clinician is able to obtain an R*1 value at a criterion level, it seems likely that the other clinicians could have also obtained that level save for their additional degree of clinician error. It should be recognized that, even for a clinician who obtains a criterion R*1 value, his ratings still consist of some portion of both types of error.

Another indication of the contribution of the two types of error to the convergence indices is in the relationship between R*1 values across clinicians in one rating condition (inter-rater reliability; test condition). As noted earlier, inter-rater reliability estimates suffer only from the first type of error whereas intra-rater estimates include both types. If the R*1 inter-rater value is high, but yet all of the R*1 intra-rater values are low, one would presume a fair measure of the second type of error (trait difference) is present.

In this regard, we see that Clinician #1 achieves a criterion R*1 value on Factors 1, 4, 6, 7, 9, 11, and 14, or, on 7 out of the 12 factors on which any clinician obtained a criterion R*1 value. Clinician #2 obtains a criterion R*1 value on Factors 4, 6, 8, 9, 10, 11, or on 6 out of the 12 factors. Clinician #3 obtains a criterion R*1 estimate on Factors 2, 6, 11, 13, 15 or on 5 out of the 12 overall factors. Mean R*1 estimates differ only slightly between clinicians: .41, .37, .38.

Factors where one or more clinicians do not achieve a criterion R*1 value, but where at least one clinician does, are, for Clinician #1

Factors 2, 8, 10, 13, 15; for Clinician #2: Factors 1, 2, 7, 13, 14, 15; and for Clinician #3: Factors 1, 4, 7, 8, 9, 10, 14.

Let us examine some areas where low R*1 estimates were obtained. From what has already been said, we see that interpretive problems were of two main types. For individual clinicians, these would be factors on which one or more clinicians did not achieve a criterion R*1 value even when a criterion R*1 estimate was obtained by at least one other clinician on that same factor. Interpretive problems for all clinicians collectively would be factors on which no clinician achieved a criterion R*1 estimate.

Interpretive Problems: All Clinicians Collectively

It was previously noted that R*1 estimates were below criterion for all three clinicians on Factors 3 (oral communication), 5 (interpersonal effectiveness), 16 (autonomy), 17 (responsibility), and 18 (overall rating). With the exception of Factors 17 and 18, all of the factors noted above are of the interpersonal, oral persuasiveness type (Appendix 1). It may be that these interpersonally oriented factors are only poorly or differentially appraised by psychometric and/or interview means, a possibility raised by Hendricks (1969). It may also be tenable that, since several tests or subtests are integrated by the clinician in rating any single factor, differences across test evaluations on the same candidate are of concern, a possibility which would explain Little & Schneidman's (1959) study.

The error inherent in the low R*1 estimates for all

clinicians on these factors might be thought of as reflecting more appraisal (type 2) error rather than clinician (type 1) error. Real differences in the levels of ability may have been present which would require that a "perfect" clinician obtain a low R*1 value in order to reflect this actual difference.

Factor 17 (responsibility) and Factor 18 (overall estimate) are the two other factors on which no clinician obtained a criterion R*1 value. The problems in interpreting these two factors are similar. What is being measured varies from condition to condition, or, in the case of Factor 18, within conditions. In the interview condition, it is logical to assume judgements of responsibility were based on past performance and quite possibly interpersonal persuasiveness. In the psychometric condition, personnel tests, which largely measure personality characteristics, were used. The differences between what is seen (interview) and what is seen to be seen (test) could account for this difference.

With Factor 18, this problem is complicated since clinicians noted that they had difficulty in separating their evaluation in terms of suitability for a particular job versus their evaluation of suitability in terms of all candidates seen. This difficulty is reflected in the low R*1 values for all clinicians, particularly Clinician #1.

Interpretive Problems: Individual Clinicians

Clinician #1

It was previously noted that Clinician #1 obtained R*1 estimates

below criterion on Factors 2 (common sense), 8 (adaptability), 10 (readiness to learn), 13 (efficiency of application), and 15 (supervisory effectiveness) even though at least one other clinician obtained criterion R*1 estimates on these factors. Since at least one clinician does obtain a criterion R*1 value, it seems likely that we are dealing with increased clinician error (type 1) on these factors. On examining these interpretively difficult factors in the light of the factor analysis already described, it is reassuring to note that they are spread over 4 of the prime 5 factors. For purposes of interpretation and evaluation then, this would appear better than if these were clustered within one prime factor rendering this factor tenuous for prediction purposes.

Once again, an examination of the raw data reveals that seldom is a candidate ranked differently than one point between rating conditions. It may be that the measure R*1 is too sensitive given the meaning and purpose of the ratings.

These factors all have a common description (Appendix 1) in that they are concerned with applied, concrete operations which may be difficult to assess in an interview setting; i.e., prediction of on-the-job applied skills.

Clinician #2

As indicated earlier, Clinician #2 obtained below criterion scores on Factors 1 (intelligence), 7 (self-reliance), 13 (efficiency of application), 14 (self-confidence), and 15 (supervisory effectiveness) even though at least one or more clinicians reached

criterion on these factors. Type 1 (clinician error) would be presumed to be higher on these factors than it would be on those factors where a criterion R*1 value was obtained.

As was the case with Clinician #1, those factors on which a low R*1 value was obtained are spread over most of the five prime factors isolated by factor analysis rather than being clustered wholly within one prime factor. Factor III (resourcefulness), however, does contain two of these low R*1 factors which might render cross-condition predictions rather tenuous. From an observation of the raw data, it is also apparent that seldom does an actual ranking difference between rating conditions exceed one for any of these factors. This further reduces the risk of actual differences in behavioral predictions based on numerically assigned differences between rating conditions. On further examining these factors in the light of the definitions given in Appendix 1, it is evident that they fall into two general areas; intellectual ability and independent self-directed work style.

Clinician #3

Clinician #3 obtained below criterion R*1 values on Factors 1 (intelligence), 4 (self-starting work drive), 7 (self-reliance), 8 (adaptability), 9 (potential for growth), 10 (readiness to learn) and 14 (self-confidence). As was the case with the other two clinicians, the low R*1 factors are spread across all of the five major factors isolated by factor analysis. with the exception of Factor II (potential ability) which includes three of them.

Commonalities

When looked at individually, it seems that clinicians had the most difficulty in rating convergently factors concerned with inner-directedness, work style and application, future potential from the perspective from learning and application, and applied intellectual problem solving.

Conclusions and Indications

1. Convergence across rating conditions is a far more elusive standard than is convergence within rating conditions. A comparison of inter-rater reliability indices with any of the intra-rater indices shows this very clearly. Logically, this is so because of the two different types of error discussed at several points.

2. With no exceptions, the most reliable indicator for purposes of analysis or prediction is the simple arithmetic mean of the three independent ratings for any given factor. In most cases, this raises the reliability index by .20 or .30.

3. On looking at the similarity between ratings across conditions (F and R*1), one can seriously question the value of the interview technique as an evaluation tool. Combined ratings, which are the ones actually used for prediction in the organization, most closely resemble those of the test ratings. Interviews are expensive and seem to contribute only inconsistency; this being aside from their obvious public relations function! This is in line with Webster's (1964) findings.

4. Differences in clinical decisions made by individual

clinicians vis-à-vis prediction cannot be discounted. It is necessary to look at, not only how well a candidate is predicted to perform, but who is making that prediction as well. If one combines the best ratings of all clinicians, the power of our "super clinician" is tremendous. If one combines the worst ratings. . .

5. Reliability, as it has come to be referred to in the literature, is not an adequate construct to use in comparing ratings across conditions. Even though we know that we have two distinct sources of error, we act as if we have only one by clinging to a traditional conceptualization.

6. Although clinicians tend to look at the same general prime areas for purposes of personnel evaluation (factor analytic interpretation), the differences are in the weighting of these factors for decision making.

7. Differences in mean ratings across conditions (high F) do not necessarily lead to differences in reliability (R*1) or vice versa. The consideration of either aspect singly is folly.

8. Factors of an interpersonal oral persuasiveness nature tend to be differentially rated by all clinicians.

9. As noted by Bray & Grant (1966), all factors are utilized for decision making but some contribute a great deal more in terms of weighting.

10. Most of the key characteristics can be evaluated by interview (Grant & Bray, 1969), but that evaluation is often diffuse and differential vis-à-vis more "objective" criteria.

11. Even though clinicians differ widely in the amount of experience they bring to this study, there is little apparent difference in level or style of decision making. This is in accord with Goldberg (1970) and Stricker (1967).

12. Sawyer (1966) may be correct in describing the main use of the interview as in providing additional, non-psychometric information to the evaluation process. However, when that information is processed psychometrically, it does not concur with other ratings of similar abilities.

13. In all cases where mean differences between rating conditions were noted, test results appeared to moderate interview impressions when rating in the combined condition.

14. Perhaps Holt (1970) was most correct of all when he said "...clinical psychologists vary considerably in their ability to do the job, but the best of them can do very well " (p. 348).

Suggestions for Further Research

1. The most major suggestion must be in the area of predictive validity. Even though we now know a great deal regarding the reliability (convergence) of clinical judgement, what is the predictive validity of judgements made in a cross-condition rating? Which rating condition best predicts future on-the-job behaviors? This would, of necessity, be a longitudinal study since only approximately 20% of the subjects involved in this study actually became employees of the companies for whom they were appraised. Because of the difficulties of comparing supervisor ratings

(criterion) with clinicians ratings of future performance, it would be hard to maintain the same degree of ecological or external validity attained in this study.

2. It would be interesting and worthwhile, if any predictive validity study were undertaken, to ascertain the factor or factors (of either the 18 by 18 matrix or the 5 by 5 matrix) which either singly or in linear combination would best predict future performance. This study would be plagued by the same external validity problems as would #1 above, but is very important research.

3. Although there are small differences between the results from each clinician, clinician sample size is too small to make any generalizable conclusions. A larger clinician sample might address itself to problems of clinical judgement, especially areas such as amount of professional training and amount of experience, problems raised by Goldberg (1970), Stricker (1967), Borke & Fiske (1957) and Oskamp (1965). It may be that as much would be lost as would be gained in this type of procedure vis-à-vis external validity. If a large number of clinicians were used, many clinicians would be called on to do tasks that they do not normally do because of experimental convenience.

4. Judgement simulation, while possible with the present data, was not a central aspect of this treatise. What is the possibility of linearly or exponentially combining single decisions in order to predict other decisions? Need we have a clinician at all or would we be better at simulating a clinician when he is at

his best or most consistent? Holt (1970) addresses these problems but not in any rigorous experimental sense.

5. How might psychological trainees be trained to simulate or duplicate the decisions of our three "experts"? Would such a procedure be viable or desirable? Stricker (1967) would see this as feasible but what would be lost by such an approach?

6. The two threats to reliability noted in this study merit examination from conceptual and practical perspectives.

7. Replication of this study (or aspects of it) on a non-industrial clientele would contribute greatly in the area of generalizability.

8. What is the generalizability of the factor analytic combination of the 18 by 18 matrix? What is the effect of feedback about your own decisions on future decisions?

9. What is the effect, in terms of actual behavioral prediction, of ratings differing only by one point? Are our statistics too powerful for our procedures?

10. What is the cost effectiveness or utility of the various approaches? What is gained by the three approaches and is it worth the price?

REFERENCES

Albrecht, P. A., Glaser, E. M., & Marks, J. Validation of a multiple assessment procedure for managerial personnel. Journal of Applied Psychology, 1964, 48, 351-360.

Ash, P., & Kroeker, L. P. Personnel selection, classification, and placement. In M. Rosenzweig & L. W. Porter (Eds.), Annual review of psychology. Palo Alto, California: Annual Reviews Inc., 1975, 481-508.

Baskett, G. B. Interview decisions as determined by competency and attitude similarity. Journal of Applied Psychology, 1973, 57, 343-345.

Bieri, J., Atkins, A., Braiai, S., Leaman, R., Miller, H., & Tripodi, T. Clinical and social judgement: the discrimination of behavioral information. New York: Wiley, 1966.

Blumenfeld, W. S. Early identification of managerial potential by means of assessment centers. Atlanta Economic Review, 1971, 21, 35-38.

Borke, H., & Fiske, D. W. Factors influencing the prediction of behavior from a diagnostic interview. Journal of Consulting Psychology, 1957, 21, 78-80.

Bray, D. W., & Grant, D. L. The assessment centre in the measurement of potential for business management. Psychological Monographs, 1966, 80 (17, Whole No. 625).

Bray, D. W., & Moses, J. L. Personnel Selection. In P. Mussen & M. Rosenzweig (Eds.), Annual review of psychology. Palo Alto, California: Annual Reviews Inc., 1972.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Campbell, J. T., Otis, J. L., Liske, R. E., & Prien, E. P. Assessments of higher level personnel: II. Validity of the over-all assessment process. Personnel Psychology, 1962, 15, 63-74.

Cronbach, L. J., & Gleser, G. C. Psychological tests and personnel decisions. Urbana: University of Illinois Press, 1965.

Dawes, R. M. Slitting the decision makers throat with Occam's Razor: the superiority of random linear models to real judges. ORI Research Bulletin, 1972, Vol. 12, No. 13.

Dawes, R. M., & Corrigan, B. Linear models in decision making. Psychological Bulletin, 1974, 81, 95-106.

Donaldson, R. J. Validation of the internal characteristics of an assessment center using the multitrait-multimethod approach. Unpublished Ph.D. thesis. Case Western Reserve University, 1969.

Dunnette, M. D. The assessment of managerial talent. In P. McReynolds (Ed), Advances in psychological measurement (Vol. 2). Palo Alto, California: Science and Behavior Books, 1971.

Einhorn, H. J. The use of nonlinear compensatory models in decision making. Psychological Bulletin, 1970, 73, 221-230.

Einhorn, H. J. Use of nonlinear, noncompensatory models as a function of task and amount of information. Organizational Behavior and Human Performance, 1971, 6, 1-27.

Einhorn, H. J. Expert measurement and mechanical combination. Organizational Behavior and Human Performance, 1972, 7, 86-106.

Erickson, E. H. The nature of clinical evidence. In D. Lerner (Ed), Evidence and inference. Glencoe, Ill: Free Press, 1959, 73-95.

Ferguson, G. A. Statistical analysis in psychology and education. Toronto: McGraw Hill, 1971.

Goldberg, L. R. Diagnosticians versus diagnostic signs: the diagnosis of psychosis versus neurosis from the MMPI. Psychological Monographs, 1965, 79 (9, Whole No. 602).

Goldberg, L. R. Reliability of peace corps selection boards: a study of interjudge agreement before and after board decisions. Journal of Applied Psychology, 1966, 50, 400-408.

Goldberg, L. R. Simple models or simple processes? Some research on clinical judgements. American Psychologist, 1968, 23, 483-496.

Goldberg, L. R. Man vs. model of man: a rationale, plus some evidence for a method of improving on clinical inferences. Psychological Bulletin, 1970, 73, 422-432.

Goldberg, L. R. Five models of clinical judgement: an empirical comparison between linear and nonlinear representations of the human inference process. Organizational Behavior and Human Performance, 1971, 6, 458-479.

Goldberg, L. R., & Werts, C. E. The reliability of clinicians' judgements: a multitrait-multimethod approach. Journal of Consulting Psychology, 1966, 30, 199-206.

Gough, H. G. Clinical versus statistical prediction in psychology. In L. Postman (Ed), Psychology in the making. New York: Knopf, 1962.

Grant, D. L., & Bray, D. W. Contributions of the interview to the assessment of management potential. Journal of Applied Psychology, 1969, 53, 24-34.

Henrichs, J. R. Comparison of "real life" assessments of management potential with situational exercises, paper-and-paper ability tests, and personality inventories. Journal of Applied Psychology, 1969, 53, 425-432.

Hoffman, P. J., Slovic, P., & Rorer, L. G. An analysis of variance model for the assessment of configural cue utilization in clinical judgement. Psychological Bulletin, 1968, 69, 338-349.

Hollman, T. D. Employment interviewers' errors in processing positive and negative information. Journal of Applied Psychology, 1972, 56, 130-134.

Holt, R. R. Clinical and statistical prediction: a reformulation and some new data. Journal of Applied and Social Psychology, 1958, 56, 1 -12.

Holt, R. R. Yet another look at clinical and statistical prediction or, is clinical psychology worthwhile? American Psychologist, 1970, 25, 337-349.

Langdale, J. A., & Weitz, J. Estimating the influence of job information on interviewer agreement. Journal of Applied Psychology, 1973, 57, 23-27.

Lipsett, L. Selecting personnel without tests. Personnel Journal, 1964, 51, 648-654.

Little, K. B., & Schneidman, E. S. Congruencies among interpretations of psychological test and anamnestic data. Psychological Monographs, 1959, 73 (6, Whole No. 476).

Mayfield, E. C. The selection interview- a reevaluation of published research. Personnel Psychology, 1964, 17, 239-260.

Mayfield, E. C., & Carlson, R. E. Selection interview decisions: first results from a long-term research project. Personnel Psychology, 1972, 25, 41-53.

McReynolds, P. An introduction to psychological assessment. In P. McReynolds (Ed.), Psychological assessment (Vol. 1). Palo Alto, California, Science and Behavior Books, 1968.

Meehl, P. Clinical versus statistical prediction. Minneapolis: University of Minnesota Press, 1954.

Michel, J. O. Assessment center validity: a longitudinal study. Journal of Applied Psychology, 1975, 60, 573-579.

Moses, J. L. The development of an assessment center for the early identification of supervisory potential. Personnel Psychology, 1973, 26, 569-580.

Oskamp, S. Overconfidence in case-study judgements. Journal of Consulting Psychology, 1965, 29, 261-265.

Perez, F. I. An experimental analysis of clinical judgement. Dissertation Abstracts International, 1973, 73-15532.

Sawyer, J. Measurement and prediction, clinical and statistical. Psychological Bulletin, 1966, 66, 178-200.

Schwab, D. P., & Heneman, H. G. Relationship between interview structure and interinterviewer reliability in an employment situation. Journal of Applied Psychology, 1969, 53, 214-217.

Shinedling, M. M., Howell, R. J., & Carlson, G. Another perspective on clinical judgement. Psychological Reports, 1975, 36, 383-389.

Slovic, P. Cue consistency and cue utilization in judgement. American Journal of Psychology, 1966, 79, 427-434.

Slovic, P., Rorer, L. G., & Hoffman, P. J. Analyzing use of diagnostic signs. Investigative Radiology, 1971, 6, 18-26.

Snow, R. E. Representative and quasi-representative designs for research on teaching. Review of Educational Research, 1974, 44, 265-292.

Spitzer, M. E., & McNamara, W. J. A managerial selection study. Personnel Psychology, 1964, 17, 19-40.

Stricker, G. Actuarial, naive clinical and sophisticated clinical prediction of pathology from figure drawings. Journal of Consulting Psychology, 1967, 31, 492-494.

Trankell, A. The psychologist as the instrument of prediction. Journal of Applied Psychology, 1959, 43, 170-175.

Ulrich, L., & Trumbo, D. The selection interview since 1949. Psychological Bulletin, 1965, 63, 100-116.

Vaughn, C. L., & Reynolds, W. A. Reliability of personal interview data. Journal of Applied Psychology, 1951, 35, 61-63.

Wainer, H. Estimating coefficients in linear models: it don't make no nevermind. Psychological Bulletin, 1976, 83, 213-217.

Webster, E. C. Decision making in the employment interview. Montreal: Eagle Publishing Company, 1964.

Wiggins, N., & Hoffman, P. J. Three models of clinical judgement. Journal of Abnormal Psychology, 1968, 73, 70-77.

Wilson, J. E., & Tatge, W. A. Assessment centers- further assessment needed. Personnel Journal, 1973, 52, 172-179.

Winer, B. J. Statistical principles in experimental design. Toronto: McGraw Hill, 1971.

Wollowick, H. B., & McNamara, W. J. Relationship of the components of an assessment center to management success. Journal of Applied Psychology, 1969, 53, 348-352.

APPENDIX 1

Definition of the 18 Characteristics Used in the Study

APPENDIX I

FACTOR DEFINITIONS

General Intelligence. Basic general capacity to learn and understand.

Readiness to Learn. The individual's willingness to acquire new information, explore new ideas, methods, tasks, etc.

Common Sense. The degree of ability to reach quick, practically-effective decisions about uncomplicated situations where sound judgement depends primarily on accumulated life and work experience, established precedent and procedures, etc.

Management-Level Planning and Problem-Solving. The individual's ability to recognize the full depth and breadth of situations and problems and to consider the longer-range, as well as the here-and-now, consequences of their change or resolution.

Oral Communication. The degree of clarity and ease with which the individual expresses himself in face-to-face discussion.

General Energy Level. The level of physical vigor and vitality the individual will demonstrate in his day-to-day conduct.

Self-Starting Work Drive. The degree to which the individual characteristically keeps himself continuously occupied in work-related activities without need of stimulation from his supervisor.

Efficiency of Application. The economic and productive organization and application of work time and effort.

General Interpersonal Effectiveness. The level of effectiveness the individual demonstrates in day-to-day dealings with others

with regard to gaining and maintaining their respect for his ideas and opinions, their confidence in his integrity, and their general feelings of good will.

Self-Confidence. The degree of basic security the individual feels in his own ability to deal adequately with more situations and people he encounters.

Leadership Force. The amount of influence and dominance the individual habitually exerts over groups and persons he encounters.

Supervisory Effectiveness. The individual's habitual effectiveness in directing, co-ordinating and controlling subordinates in standard work settings.

Self-Reliance. The degree to which the individual carries out assigned responsibilities without seeking direction, help, encouragement and/or reassurance from co-workers.

Autonomy. The degree of the individual's need to make his own decisions, regulate his own behavior, be his own boss, etc.

Adaptibility. The level of ability to cope comfortably with new and changing circumstances.

Responsibility. The degree to which the individual lives up to personal, professional and business obligations he has tacitly or otherwise accepted.

Potential for Growth. The degree of probability that the individual will develop the personal resources to cope with increasingly

more complex and responsible work roles.

General Suitability for Job Concerned. Self-explanatory.

APPENDIX 2

Interview Rating Form

CANDIDATE'S NAME _____

CANDIDATE'S AGE _____

ASSIGNMENT NUMBER _____ NAME _____

DATE _____

RATER'S NAME _____

Rate the candidate on each of the following characteristics according to the following code. Place the number that represents the most correct description in the space provided opposite each characteristic.

1 = Poor; 2 = Marginal; 3 = Adequate; 4 = Good; 5 = Very Good

If you are genuinely unable to rate a candidate on a characteristic, leave the space opposite that characteristic blank.

1. General Intelligence	____	10. Readiness to Learn	____
2. Common Sense	____	11. Management Level Planning	____
3. Oral Communication	____	12. General Energy Level	____
4. Work Drive	____	13. Efficiency of Application	____
5. Interpersonal Effect.	____	14. Self-Confidence	____
6. Leadership Force	____	15. Supervisory Effectiveness	____
7. Self-Reliance	____	16. Autonomy	____
8. Adaptability	____	17. Responsibility	____
9. Potential for Growth	____	18. General Suitability	____

INTERVIEW RATING FORM

APPENDIX 3

Interview + Test Rating Form

CANDIDATE'S NAME _____

CANDIDATE'S AGE _____

ASSIGNMENT NUMBER _____ NAME _____

DATE _____

RATER'S NAME _____

Rate the candidate on each of the following characteristics according to the following code. Place the number that represents the most correct description in the space provided opposite each characteristic.

1 = Poor; 2 = Marginal; 3 = Adequate; 4 = Good; 5 = Very Good

If you are genuinely unable to rate a candidate on a characteristic, leave the space opposite that characteristic blank.

1. General Intelligence	____	10. Readiness to Learn	____
2. Common Sense	____	11. Management Level Planning	____
3. Oral Communication	____	12. General Energy Level	____
4. Work Drive	____	13. Efficiency of Application	____
5. Interpersonal Effect.	____	14. Self-Confidence	____
6. Leadership Force	____	15. Supervisory Effectiveness	____
7. Self-Reliance	____	16. Autonomy	____
8. Adaptability	____	17. Responsibility	____
9. Potential for Growth	____	18. General Suitability	____

INTERVIEW + TEST RATING FORM

APPENDIX 4

Test Rating Form

CANDIDATE'S NAME _____

CANDIDATE'S AGE _____

ASSIGNMENT NUMBER _____ NAME _____

DATE _____

RATER'S NAME _____

Rate the candidate on each of the following characteristics according to the following code. Place the number that represents the most correct description in the space provided opposite each characteristic.

1 = Poor; 2 = Marginal; 3 = Adequate; 4 = Good; 5 = Very Good

If you are genuinely unable to rate a candidate on a characteristic, leave the space opposite that characteristic blank.

1. General Intelligence	____	10. Readiness to Learn	____
2. Common Sense	____	11. Management Level Planning	____
3. Oral Communication	____	12. General Energy Level	____
4. Work Drive	____	13. Efficiency of Application	____
5. Interpersonal Effect.	____	14. Self-Confidence	____
6. Leadership Force	____	15. Supervisory Effectiveness	____
7. Self-Reliance	____	16. Autonomy	____
8. Adaptability	____	17. Responsibility	____
9. Potential for Growth	____	18. General Suitability	____

TEST RATING FORM

APPENDIX 5

Dunnette (1971) Table 1

Table 1
Assessment Methods Showing High Correlations with Each of Eight Behavior Rating Factors and Overall Staff Prediction for College and Non-College Men in the AT&T Management Progress Study

Assessment method	College men	Non-college men
Factor I. General Effectiveness		
Performance in Cooperative Group Exercise		.60
Performance in Competitive Group Exercise	.67	
Performance on In-Basket	.60	.59
Interview: Personal Impact	.52	.48
Projective: Leadership Role	.48	.51
Personality Test: Dominance	.33	.22
Factor II. Administrative Skills		
Performance on In-Basket	.76	.68
Performance in Competitive Group Exercise	.48	.51
Mental Ability Test	.34	.72
Interview: Personal Impact	.42	.24
Oral Communications Skills	.33	.53
Projective: Leadership Role	.36	.36
Personality Test: Dominance	.30	.30
Factor III. Interpersonal Skills		
Performance in Cooperative Group Exercise	.39	.52
Performance in Competitive Group Exercise	.62	.45
Performance on In-Basket	.45	.49
Interview: Personal Impact	.44	.25
Human Relations Skills	.28	.46
Factor IV. Control of Feelings		
Performance in Competitive Group Exercise	.47	.36
Performance in Cooperative Group Exercise	.37	.35
Interview: Human Relations Skills	.23	.45
Tolerance of Uncertainty	.30	.40
Projective: Leadership Role	.29	.46
Dependence	−.28	−.42
Factor V. Intellectual Ability		
Mental Ability Test	.70	.62
Interview: Oral Communications Skills	.40	.47
Factor VI. Work Orientation Motivation		
Projective: Work or Career Orientation	.50	.56
Interview: Personal Impact	.36	.50
Inner Work Standards	.40	.43
Performance in Cooperative Exercise	.30	.39
Performance in Competitive Exercise	.45	.36
Performance on In-Basket	.44	.26
Factor VII. Passivity		
Interview: Need Advancement	−.57	−.67
Personal Impact	−.38	−.58
Need Security	.50	.37
Projective: Leadership Role	−.47	−.40
Achievement Motivation	−.41	−.50
Performance in Competitive Exercise	−.39	−.36
Performance in Cooperative Exercise	−.35	−.34
Personality Test: General Activity	−.43	
Factor VIII. Dependency		
Projective: Affiliation	.46	.41
Dependence	.49	.37
Overall Staff Prediction		
Performance in Competitive Exercise	.60	.38
Performance on In-Basket	.55	.51
Performance in Cooperative Exercise	.41	.42
Interview: Personal Impact	.49	.21
Oral Communications Skills	.41	.48
Projective: Achievement Motivation	.30	.40

APPENDIX 6

Principal Components Factor Analysis (Varimax Rotation) of Characteristics Appraised in Test Rating Condition

21

APPENDIX 6a

Principal Components Factor Analysis (Varimax Rotation)

of 18 Characteristics Appraised in Test Rating Condition:

Clinician #1 (N=74)

Appraised Characteristic	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	H*2
1	.83	-.09	-.05	.07	.09	.70
2	.22	.16	.11	.05	-.61	.46
3	.38	.77	-.13	.18	.06	.79
4	-.10	.10	-.03	.83	-.03	.71
5	.24	.73	.26	-.08	-.27	.74
6	-.17	.76	.19	.09	.26	.71
7	-.04	.20	.38	.19	.62	.60
8	.77	.11	.16	-.20	-.09	.68
9	.84	.29	.18	.10	.05	.83
10	.82	.13	.20	-.13	-.20	.78
11	.86	.06	-.06	.80	-.003	.75
12	.12	.05	.39	.76	.17	.77
13	.02	-.03	.85	.06	.07	.74
14	.05	.80	-.03	.24	.12	.72
15	.12	.72	.32	-.33	.01	.75
16	.20	.23	-.04	.03	.73	.62
17	.10	.20	.84	.12	-.14	.78
18	.47	.41	.63	.001	.12	.80
Variance	3.96	3.32	2.45	1.61	1.56	12.91
% of Total Variance	22%	18%	14%	9%	9%	72%

APPENDIX 6b

Principal Components Factor Analysis (Varimax Rotation)

of 18 Characteristics Appraised in Test Rating Condition:

Clinician #2 (N=74)

Appraised Characteristic	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	H*2
1	.76	.02	-.03	.14	.32	.70
2	.25	.22	-.23	.68	.08	.64
3	.82	-.02	.04	.21	-.07	.72
4	-.002	.05	.14	.04	.83	.71
5	.17	-.09	.25	.71	.01	.60
6	-.27	.46	.38	.36	-.12	.57
7	.04	.25	.80	-.06	-.005	.70
8	.66	.29	.11	.05	-.37	.67
9	.71	.45	.13	.14	-.13	.76
10	.71	.41	-.08	-.05	-.29	.76
11	.85	-.03	.05	.08	.05	.73
12	.12	.68	.19	.07	.09	.53
13	-.20	.56	-.20	-.13	.53	.69
14	-.008	-.04	.75	.19	.04	.61
15	.05	.36	.21	.70	-.08	.67
16	.11	.009	.79	.07	.05	.65
17	.25	.76	-.10	.18	-.03	.69
18	.41	.65	.30	.32	.09	.79
Variance	3.87	2.70	2.37	1.89	1.36	12.18
% of Total Variance	22%	15%	13%	11%	8%	68%

APPENDIX 6c

Principal Components Factor Analysis (Varimax Rotation)

of 17 Characteristics Appraised in Test Rating Condition:

Clinician #3 (N=74)

Appraised Characteristic	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	H*2
1	-.07	.24	.12	.66	.14	.53
2	-.10	-.08	-.04	.69	-.21	.54
4	.26	-.13	.10	-.10	-.68	.57
5	.48	.26	.30	.17	-.45	.62
6	.78	-.21	.18	-.08	-.26	.76
7	.13	-.04	.75	-.17	.10	.63
8	.06	.77	-.09	-.05	.10	.61
9	-.04	.76	.31	.18	-.18	.73
10	.05	.60	.44	-.10	-.08	.57
11	.14	.65	.02	.34	.17	.58
12	.66	.01	.23	-.15	.04	.51
13	.12	.05	.59	.32	.24	.53
14	.66	.37	-.17	-.15	-.002	.63
15	.57	.09	-.14	.41	.37	.65
16	.20	-.08	.32	-.17	.61	.55
17	-.04	.25	.65	.67	-.26	.56
18	.48	.24	.51	.31	-.13	.66
Variance	2.44	2.40	2.24	1.59	1.55	10.22
% of Total Variance	14%	14%	13%	9%	9%	60%

APPENDIX 7

CAPSULE SUMMARY OF TESTS

APPENDIX 7

CAPSULE SUMMARY OF TESTS

Listed below is a short description of each of the tests used in this study. For complete information regarding a specific test, the reader is referred to the appropriate test manual.

Differential Aptitude Tests

Verbal Reasoning. This is a verbal concept understanding test. It is designed to evaluate the ability to abstract, generalize, and to think constructively. Testing format involves verbal analogies.

Abstract Reasoning. This is a non-verbal reasoning ability test. The testee is required to formulate operating principles in changing abstract diagrams. Operating principles involve the use of logic.

Wonderlic Personnel Test

The Wonderlic is a test of mental ability. It is widely used as a selection tool in hiring and as an indicator of future development possibility.

Watson-Glaser Critical Thinking Appraisal

This test involves the appraisal of important critical thinking skills (inference, recognition of assumptions, deduction, interpretation and evaluation of arguments) in everyday situations.

Business Judgement Test

This test is designed to measure empathy and knowledge of generally accepted ways of behaving in business interpersonal situations.

Test of Practical Judgement

This test is designed to evaluate the testee's ability to select the best solution to factual and complex interpersonal business problems.

Supervisory Practices Test

This test is designed to appraise supervisory ability or potential ability. It is directly concerned with supervisory thinking, attitudes, and opinions.

Management Aptitude Inventory

This inventory is designed to assess characteristics related to success in managerial positions (intelligent job performance, leadership qualities, proper job attitude, and relations with others).

Vocational Preference Inventory

This is personality questionnaire which uses preference for vocational titles as a measure of personality style. It is designed to assess areas such as interpersonal relations, interests, values, self-conception, coping behavior, and identification.

Edwards Personal Preference Schedule

This personality test provides a convenient measure of normal personality variables such as achievement, deference, order, exhibition, autonomy, affiliation, intraception, succorance, dominance, abasement, nurturance, change, endurance, heterosexuality, and aggression.

California Psychological Inventory

This is a multiple choice personality test which measures 18 personality variables in four general areas (measures of poise, ascendancy, and self assurance; measures of socialization, maturity and responsibility; measures of intellectual potential; measures of personal orientation and values).